InPharmix Incorporated

PDQ_MED

graphInPharmix

Check_Queries

Namer

StAT

Installation Instructions
Users Guides

Table of Contents

1. Introduction

Thank you for purchasing InPharmix Software. We have made every effort to provide our customers with robust software tools for biomedical research. If you have any questions, encounter any problems, or have suggestions, please contact us.

• Support

You may contact InPharmix Inc. using any of the methods listed below.

On the web:

http://www.inpharmix.com/support.htm

Electronic mail:

Support Information: Support@InPharmix.com
General Information: Info@InPharmix.com
Sales Information: Sales@InPharmix.com

Telephone:

1-(317)-422-1464

Post:

InPharmix Inc.
PO Box 406
Greenwood, IN 46142
USA

• Versions

StAT	Version 1.02 24 September 2003
PDQ_MED	Version 2.81 19 September 2003
graphInPharmix (included with **PDQ_MED**)	6 October 2003 (applet Version 3.10e)
Check_Queries	Version 0.18 11 January 2003
Namer	Version 0.24 (alpha) 7 March 2003
Documentation	26 September 2003

Disclaimer and Limitation of Liability

While we use reasonable efforts to insure the use ability of this software, we make no representations as to the accuracy, quality, timeliness, availability, or completeness of the information, software, products, or other materials. We provide this software on an "as is" basis. You use it at your own risk, and InPharmix Incorporated, its employees, distributors, directors, and agents are not liable for any errors or omissions in its content or delivery, or for any form of loss or damage (including any consequential, indirect, incidental, special, or exemplary damages, even if known to us) that may result from its use. We expressly disclaim all warranties, including warranties of merchantability, fitness for a particular purpose, or non-infringement. No warranty not set forth in this agreement will be valid. If any of the above provisions are void under governing law, our liability shall be limited to the extent permitted by law.

InPharmix Software

This software is proprietary and confidential information that is protected by applicable intellectual property and other laws. Except as expressly authorized by us, you will not modify, sell, or distribute the software. We grant you a non-transferable and non-exclusive license to use the software on a single computer; provided that you do not (and do not allow any third party to) copy, modify, reverse engineer, assign or otherwise transfer any right in the software.

2. Installation Instructions

• Unix / Linux

Install the Files

Create a directory in your cgi directory for the top level of the InPharmix installation, for example `/usr/web/cgi/InPharmix/`. Change the protections for this directory to ug+rx;

```
chmod ug+rx /usr/web/cgi/InPharmix
```

Set default (cd) into this top-level directory. Load the distribution CD and browse to the CD's `Unix_Linux` directory. Copy the file `distribution_YOURSITE.tar` to the directory you created above. Unpack the tar file with;

```
tar xvfp distribution_YOURSITE.tar
```

You should now have the following directory structure;

```
drwxr-xr-x  ../InPharmix/
drwxr-xr-x     PDQ_MED/
drwxrwxrwx        PDQ_MED/StAT_user_data/
drwxrwxrwx        PDQ_MED/PDQMED_user_data/
drwxr-xr-x        PDQ_MED/Java/
drwxr-xr-x     StAT/
drwxrwxrwx        StAT/StAT_user_data/
drwxr-xr-x     images/
```

Check the protections on these directories and correct as needed. You may delete the `distribution_YOURSITE.tar` file.

Perl 5.005

StAT and PDQ_MED both require Perl version 5.005 or better. If needed, Perl 5.005 can be downloaded from;

```
http://www.perl.com/pub/language/info/software.html
http://www.perl.com/pub/language/info/software.html#unix
```

The standard Perl 5.005+ download should contain a module called LWP. If, when trying to run PDQ_MED or StAT, you get an error message saying the LWP module is not installed you will need to download and install it from www.perl.com. Usually, the LWP module is in a directory like `/usr/local/perl/bin/`.

Perl Path

All of the InPharmix perl programs expect the directory containing your perl executable to be at `/usr/bin/perl/`. If your perl directory is different (e.g., `/usr/local/bin/perl/`) then the InPharmix perl scripts will need to be changed. Change your working directory to

../InPharmix/ and issue the following commands to change the path to perl (this assumes you are changing the path in the perl files to the perl directory at /usr/local/bin/perl);

```
perl -p -i.bak -e 's|^\#\!\/usr\/bin\/perl|\#\!\/usr\/local\/bin\/perl|' *.pl
cd StAT
perl -p -i.bak -e 's|^\#\!\/usr\/bin\/perl|\#\!\/usr\/local\/bin\/perl|' *.pl
cd ../PDQ_MED
perl -p -i.bak -e 's|^\#\!\/usr\/bin\/perl|\#\!\/usr\/local\/bin\/perl|' *.pl
```

Perl CGI

Your server software needs to recognize files with the .pl extension as Perl executable files. Your system administrator or web guru should be able to do this. Usually, the Perl directory is either `/usr/bin/Perl` or `/usr/local/bin/Perl`.

Notes About the Apache Web Server

If your server uses the Apache Server software please note that there are three ways to configure the server for Perl cgi files; *CGI*, *PerlRun* and *Registry*. *PerlRun* and *Registry* are a part of Apache::mod_perl. We recommend using either *CGI* or *PerlRun* and not the *Registry* mode for the InPharmix directories. Your system administrator or web guru should be able to assist you in setting this up.

A Typical Apache Configuration

The Apache configuration file is typically located at a directory such as `/httpd/conf/httpd.conf` or `/usr/httpd/conf/httpd.conf`. One example of how to configure the Apache server for the InPharmix directory is;

```
Alias /InPharmix/ "/MACHINE_NAME/FULL_PATH/InPharmix/"
#use http://MACHINE_NAME/InPharmix/

<Directory /MACHINE_NAME/FULL_PATH/InPharmix>
   Options +ExecCGI +FollowSymLinks
   AddHandler cgi-script .pl
#  RLimitMEM in bytes, 200000000 ~= 200MB
   RLimitMEM 200000000
#  Rlimit CPU in seconds, 6000 = 100 minutes
   RLimitCPU 6000
#  no space in "allow,deny"
   Order allow,deny
   Allow from all
</Directory>
```

Where "`/MACHINE_NAME/FULL_PATH/`" is the complete path to the machine and directory that contains the InPharmix directory.

Server Limits

Both **StAT** and **PDQ_MED** place a moderate load on their server. For best results, we suggest that the server software be set to allow cgi scripts at least 100MB of memory and ten minutes of CPU time. In a typical Apache httpd.conf file the commands are;

```
RLimitMEM 100000000  # in bytes, 200000000 ~=200MB
RLimitCPU 6000       # in seconds, 6000 = 100 minutes
```

To test your servers limits we have included two short programs for testing time and memory limitations; `cgi_memory_limit.pl` and `cgi_time_limit.pl` (both are located in the top level InPharmix directory).

If your server is limiting time and/or memory, ask your system administrator to increase these limits. Note that you must have root privileges to change the RLimitMEM and RLimitCPU values.

To complete the installation continue at the section for "All Platforms".

• Windows

Two possible web servers for a Windows machine are the Apache Server and the Microsoft Personal Web Server (PWS).

Using the Apache Server with Windows

For more information and to download the executable code for Apache visit the Apache web site at http://www.apache.org/. If you use the Apache Server then see the section on "A Typical Apache Configuration" on suggestions for the configuration file.

Using the Microsoft Personal Web Server (PWS)

If the PC that you will be installing on is already set up as a web server you can skip this step.

If you have a Windows 98, 98SE or NT compact disk:

1. Insert your compact disc in its drive.
2. Click Start, and then click Run.
3. In Open, type:
   `x:\add-ons\pws\setup.exe`
   where x is the letter of your CD-ROM drive.
4. Click OK.
5. Follow the directions in the Personal Web Server Setup.

If you do not have the CD:

1. Click Start, then Find, then Files or Folders and search your hard drive for a directory called pws. It may be in a directory such as c:\windows\options\cabs\pws\.
2. Within the pws folder, double click on file `setup.exe` to start the installation program for pws.
3. Follow the directions in the Personal Web Server Setup.

Install the Files

Create a folder in `c:\Inetpub\wwwroot` (or whatever the directory the PWS is using) to hold the InPharmix files, for example `c:\Inetpub\wwwroot\InPharmix\`.

Load the InPharmix distribution CD and browse to the `Windows_Plain` directory.

Drag the contents of the Windows_Plain folder into the `c:\Inetpub\wwwroot\-InPharmix` folder.

You should now have the following directory structure;

```
c:\Inetpub\wwwroot\InPharmix\
c:\Inetpub\wwwroot\InPharmix\PDQ_MED\
c:\Inetpub\wwwroot\InPharmix\PDQ_MED\StAT_user_data\
c:\Inetpub\wwwroot\InPharmix\PDQ_MED\PDQMED_user_data\
c:\Inetpub\wwwroot\InPharmix\PDQ_MED\Java\
c:\Inetpub\wwwroot\InPharmix\StAT\
c:\Inetpub\wwwroot\InPharmix\StAT\StAT_user_data\
c:\Inetpub\wwwroot\InPharmix\images\
```

Set the File and Folder Protections

Check, and if necessary modify, the protections for the InPharmix folder:

1. Browse to the wwwroot folder (probably in c:\Inetpub).
2. Right click on the InPharmix folder
3. Select Properties
4. Select the Web Sharing tab
5. Select /InPharmix
6. Select Edit Properties
7. Check the Read, Execute and Scripts check boxes.
8. Click OK, then Apply.
9. Restart the PC.

Perl 5.005

StAT and PDQ_MED both require Perl version 5.005 or better. If needed, Perl 5.005 can be downloaded from;

`http://www.perl.com/pub/language/info/software.html`
`http://www.perl.com/pub/language/info/software.html#win32`

The standard Perl 5.005+ download should contain a module called LWP. If, when trying to run **PDQ_MED** or **StAT**, you get an error message saying the LWP module is not installed you will need to download and install it from www.perl.com. Usually, the LWP module is in a directory like `C:\Perl\site\lib\`.

Configure the Microsoft Personal Web Server for Perl

For additional information see; `http://support.microsoft.com/support/kb/articles/Q231/9/98.ASP`.

1. On the Start menu, click Run.
2. In the Open box, type Regedit and click OK.
3. Open the following registry key:
   HKEY_LOCAL_MACHINE
   \SYSTEM \CurrentControlSet \Services \W3SVC \Parameters \ScriptMap
4. On the Edit menu, point to New, and click String Value.
5. Name the value .pl and press ENTER.
6. Select .pl, and click Modify on the Edit menu.
7. In the Value Data box, type <the full path to perl.exe>\perl.exe %s %s
   NOTE: The "%s %s" is case sensitive (for example, "%S %S" will not work). The full path to perl.exe is usually something like c:\Perl\bin\perl.exe.
8. Close the Registry Editor, and restart your computer.

Your installation includes a "default.htm" file, which contains links to the installed software. The urls for this file is; `http://HOST_NAME/InPharmix/default.htm`

If you are using PWS then HOST_NAME is the name PWS has given your PC. If you are not sure what the name is, open the Personal Web Manager and your HOST_NAME is in the

address for your home page, i.e., http://HOST_NAME. If you are using the Apache Server, then HOST_NAME is defined in your httpd.conf file.

• All Platforms

Once the platform specific part of the installation is complete, a few changes need to be made to the configuration file `site.data` to configure the installation for your site.

The site.data & default.html Files

Your installations `site.data` file is located in the top level InPharmix directory. This file contains various defaults, a local contact person, information about your license and other information. This file must be present and valid or PDQ_MED and StAT will not run properly. The `site.data` file is a plain text file and can be edited. We suggest you change the "`licenseSite`", "`licenseScientist`" and "`licenseEmailTo`" fields to values appropriate for your site. You should also set the "opSystem" option as needed. If you have a local MEDLINE database then the options listed under "Local DB Options" will also need to be changed. If you do not have a local MEDLINE database, then all of the options starting with "`localDB_`" need to be commented out.

```
#  Site Data file
#     comments begin with a #
#     delimiter is :\t+
#     blank lines are ignored
#     NO BLANKS AT BEGINNING OF LINES
#
#  InPharmix Inc.
#
#  License file for:
#  created:          11 Sept 2003
#
licenseSite:         InPharmix Inc.      # Name of the licensee
licenseScientist:    J. Sluka            # The contact person at the licensee site
licenseEmailTo:      JSluka@InPharmix.com # local email address for licenseScientist
licenseType:         full                # possible values are "full" and "demo"
licenseExpires:      31 Dec 2003         # date the license expires, dD MON YEAR
licenseMessage:      This is a full license for InPharmix Inc. software # as approriate
#
siteMessage:                             # no message!
#
###############################################################################
#
# Define the operating system that PDQ is running on
# Just comment out the ones that are not correct
#
#opSystem:           unix      # Unix / Linux
opSystem:            win98     # Windows 98
#opSystem:           winNT     # Windows NT
#opSystem:           winXP     # Windows XP
#
###############################################################################
# Local DB Options
# July 2003  Options for using a local MEDLINE database instead of the NCBI site
# Comment out all of the lines below if a local DB is NOT and option
# PDQ_MED uses the Perl DBI module to access local databases. Any DBI
# compliant database, such as MySQL or Oracle, should work.
#
localDB_Type:        mysql               # mysql, Oracle, XBase, mSQL
#                                        #    (all case specific)
localDB_Address:     localhost           # path to the DB, or
#                                        #    host=wfudb.wfu.edu;sid=WFUD
localDB_Database:    localMEDLINEtest    # the name of the DB
```

```
localDB_Table:      MEDLINE_abs         # name of the table that includes the
#                                       #    Abstract, Title and PMIDs
localDB_TitleF:     TITLE               # name of the column in localDB_Table that
#                                       #    contains the article title
localDB_AbsF:       ABSTRACT            # name of the column in localDB_Table that
#                                       #    contains the article abstract
localDB_PmidF:      PM_ID               # name of the column in localDB_Table that
#                                       #    contains the article PMID
localDB_asDefault:  no                  # "yes" or "no", should using the localDB be
#                                       #    the default option
# Username and Password options:
localDB_usePasswd:  yes                 # "yes" for either username/passwd in this
#                                       #   file or on the web page
#                                       #   "no" if no username/passwd required
# If username and password are not needed, or if they will be input from the
# web page,  comment out these two lines
localDB_Username:   jsluka              # username
localDB_Pass:       mysql               # passwrd
#
# End Local DB Options
#################################################################################
```

In addition to the changes in the site.data file, you should also edit the default.html file (located in the same directory). In the default.html file locate the line similar to;

```
<li>Your local contact person is <a
  href="mailto:jsluka@inpharmix.com?subject=InPharmix%20Software%20
  at%20InPharmix">Dr. James Sluka</a>
```

and change the name and email address.

• Testing and Running the Software

Your installation includes a "default.html" file that contains links to the installed software. The url for this default is simply the top-level directory you created above (`http://MACHINE_NAME/InPharmix/`).

If installed, the **StAT** url will be;

`http://MACHINE_NAME/InPharmix/StAT/StAT.pl`

If installed, the **PDQ_MED** url will be;

`http://MACHINE_NAME/InPharmix/PDQ_MED/PDQ_MED.pl.`

We suggest you test the links from the default page and test the software with the default input values provided on the respective web pages.

• Web Browsers

Due to problems in the first two releases of Netscape Navigator 6, we recommend using either version 4.x or 6.2.1 or 7.x. Any version of Microsoft's Internet Explorer is acceptable.

There is a bug in Netscape Navigator 7.1 that corrupts the "Resubmit Graph" hyperlink in saved HTML files. To correct this, open the saved HTML file in a text editor (not an HTML editor) and locate the line(s) containing;

```
<form action="graphJPS_Interface.pl" ...
```

Change the "action" parameter to;

```
<form action="http://path/InPharmix/PDQ_MED/graphJPS_Interface.pl"...
```

Where **path** is the same as the path used to run PDQ_MED from a browser.

• Proxy Servers

If you site uses a proxy server, you must make sure that the InPharmix routines can local the proxy correctly. If you use the Apache server then the proxy path can be included in the httpd.conf file where the InPharmix directory is defined. For example;

```
# InPharmix directory
#
# path to the top level InPharmix directory
Alias /InPharmix/ "/export/home/InPharmix/"
<Directory /export/home/InPharmix>
  SetEnv http_proxy http://proxy.YOURSITE.com:5150/
  Options +ExecCGI +FollowSymLinks
  AddHandler cgi-script .pl
  # RLimitMEM should be >100MB
  RLimitMEM 200000000
  # RLimitCPU should be >10 minutes (600 seconds)
  RLimitCPU 6000
  Order allow,deny
  Allow from all
</Directory>
```

You could also the PasEnv directive in the httpd.cong file.

• Using a Local Database Instead of NLM/NCBI

PDQ_MED supports searching a local MEDLINE database in place of web searches at NLM/NCBI. Supported databases are mySql and Oracle. The MEDLINE data must be contained in a single database table and must include the PMID, TITLE and ABSTRACT fields.

Search results will be somewhat different using a local database compared to the NLM/NCBI web site. In particular, the local database will not use the MeSH system or the phrase index that the NLM/NCBI uses. In general, this should not be a problem and in some cases the local database may actually do a better job of searching then does NLM/NCBI.

To configure PDQ_MED to use a local database, the site.data file must be modified to include information on the type of database, its name, location and the names of the database table and columns. See the section on the site.data file for further information or contact InPharmix for help.

• Running PDQ_MED From the Command Line

It is possible to run PDQ_MED from the command line instead of via a web browser. Because of the large number of input parameters that PDQ_MED uses, it is easiest to create a "batch" file to control the PDQ_MED job.

As an example, we will create a csh batch file for use on Unix/Linux platforms. The batch file will do a proximity search, skip the pharma search, using the terms raloxifene, estrogen, and "Acute Myeloid Leukemia".

```
#!/bin/csh
# ========================================================
#FOR CSH
#   no spaces around equals (abc=123 NOT abc = 123):
```

```
#   use \ for newline (in queries and optionally between parameters)
#   use single quotes around the queries to protect the spaces
#   and double quotes in MEDquery
# =======================================================
perl PDQ_MED.pl \
MEDquery='raloxifene\
estrogen\
[[disease]] "Acute Myeloid Leukemia" "Acute Myelogenous Leukemia"'\
Proximity=yes\
field=TIAB\
StateToState=yes\
ENRtoENR=no\
Proximity=yes\
Pharma=no\
PharmaTerms=''\
GroupingType=raw\
GroupingCutOff=1\
DisplayGraph=yes\
MaxIndex=5000\
MaxAbs=50  > MYOUTPUT_FILE.html
# =======================================================
```

Note the optional redirection of the output at the end of the batch file.

The complete set of queries must be entered inside of single quotes ('), one query per line with lines terminated with a backslash. Use double quotes (") for quotes within the set of queries.

Complete list of PDQ_MED input Options

Note that the parameter names are case specific.

Parameter (case specific)	Description	Range	Default
project	name for the project	text	
Infile	**CANNOT BE USED IN A BATCH FILE**		
MEDquery	set of queries	text	none
field	MEDLINE field to be searched	TI, TIAB, ABS, ALL…	TIAB (title + abstracts)
StateToState	search [[state]] and [[disease]] terms against each other	yes no	yes
ENRtoENR	search ENR terms against each other	yes no	no
Proximity	turn on proximity searching	yes no	no
Pharma	turn on pharma searching	yes no	no
PharmaTerms	set of pharma queries	text	same as PDQ_MED
GroupingType		raw weighted	raw
GroupingCutOff	minimum number of co-occurences that trigger grouping	> 0	1
DisplayGraph	turn display of the graphs (Java applets) on and off	yes no	yes
MaxIndex	maximum number of PMIDs for indexing per term	> 1	20,000
MaxAbs	maximum number of abstracts to check per term pair	> 1	250
GlobalTerm	global search term	text	none
LimitPair-WiseSearches	only do proximity searches for the first N terms	> 1	none
useLocalDB	toggle between a local database and NCBI/NLM for the searches	yes no	no

localDB_Type	type of local database (**case specific**)	mysql Oracle	
localDB_Address	path to the DB. e.g., "localhost" or something like "host=wfudb.wfu.edu;-sid=WFUD"		
localDB_Database	name of the DB		
localDB_Table	DB table that includes the Abstract, Title and PMIDs		
localDB_TitleF	column in localDB_Table that contains the article title		
localDB_AbsF	column in localDB_Table that contains the article abstract		
localDB_PmidF	column in localDB_Table that contains the article PMID		
localDB_Username	username for DB access		
localDB_Pass	password for DB access		

3. PDQ_MED Users Guide

• Introduction:

PDQ_MED (Pair-wise Data Query to MEDLINE) analyzes the MEDLINE biomedical literature concerning a group of terms. PDQ_MED is particularly useful for discovering connections in the literature for a set of genes and/or proteins that have been linked under some experimental paradigm, such as from a gene chip, tissue library, or subtracted library experiment.

PDQ_MED is unique in that it is NOT statistically or lexically based like most text mining tools. Instead, PDQ_MED searches for and analyses abstracts, which interconnect two or more of the user's, query terms. This method insures that even rare connections in the literature, such as from seemingly unrelated areas of research, are identified. Most statistically based text analysis methods will ignore infrequent connections.

As an example, consider the query set;

```
HGH "human growth hormone"
estradiol estrogen oestrogen 17b-estradiol
Interleukin-1b Interleukin-1beta IL-1b
```

Will, in effect, launch the following queries to MEDLINE;

```
(HGH OR "human growth hormone")
(HGH OR "human growth hormone") AND (estradiol OR estrogen
   OR oestrogen OR 17b-estradiol)
(HGH OR "human growth hormone") AND (Interleukin-1b
   OR Interleukin-1beta  OR IL-1b)

(estradiol OR estrogen OR oestrogen OR 17b-estradiol)
(estradiol OR estrogen OR oestrogen OR 17b-estradiol)
   and (Interleukin-1b OR Interleukin-1beta OR IL-1b)

(Interleukin-1b OR Interleukin-1beta OR IL-1b)
```

The numbers of abstracts that are found by the searches listed above are the basis for discovering linkages between your query terms. After all possible pair-wise term searches have been completed, PDQ_MED uses a greedy clustering algorithm to group the terms. If term-A and term-B co-occur in a set of abstracts and term-B and term-C co-occur in another set of abstracts, then term-A, term-B and term-C are all members of the same group.

In addition, PDQ_MED allows the user to do "proximity searching". In a proximity search, a pair of terms is not only required to co-occur in the same abstract but must also co-occur in the same sentence.

NOTE: Currently the **sentence proximity** and **Pharma Terms** checking is limited to the first 500 abstracts returned for a given query. MEDLINE usually returns the most recent abstracts first so the analysis is limited to the 500 most recent abstracts for a given query pair.

• New Features in Version 2.8

Improved Search Algorithm

The search algorithm has been changed somewhat in version 2.8 of PDQ_MED. In particular, PDQ_MED now searches for plurals and possessives of queries when doing proximity searching. For example, the query "catalase" now matches "catalases" and "catalase's".

Local Database Option

PDQ_MED now supports searching a local MEDLINE database in place of web searches of the NLM/NCBI Entrez database.

Improvements in the Graph Display

The graph algorithm has been changed and now provides a more stable graphic display. In addition, the graph applet now tries to minimize edge crossings and stops automatically after 1001 iterations.

New Output Format Option

A new output report format option has been added. In addition to the "Groups" report format, there is a new "Lists" report format. The Lists report format focuses on relationships between the query genes and a "master" disease/state term.

Query Rate

It is possible to overload the MEDLINE/Entrez web site if your site has a fast internet connection and you are running several PDQ_MED jobs. PDQ_MED now tracks the total number of PDQ_MED jobs running on the server. As the number of jobs running increases, PDQ_MED slows down the rate at which queries are sent to the NCBI server.

• New Features in Version 2.5

Global Search Term

An option to include a "Global Term" in all of the searches has been added. Pairs of terms are searched in MEDLINE as Term-1 AND Term-2 AND GlobalTerm.

Limit Pairwise Searches

An option to only do pair-wise searches for the first N terms vs. all of the remaining terms has been added.

Additional Search Options

Options to not search disease/state terms and "Expressed but not regulated" (ENR) genes against each other has been added. The option to only do a Pharma Search has also been added.

Default MEDLINE Search Field

The default MEDLINE search field has been changed from "All" to "Title/Abstract".

Additional Output Files

Additional output files are now created for disease and state terms.

Changes To The Output

The output has been streamlined by moving some of the tables to external files. In particular, the "Insignificant" cross references have been moved to an external file. An "Insignificant" cross-reference is a co-occurence of a pair of terms in an abstract that failed proximity checking.

Query Rate

A time delay has been introduced to reduce the rate at which queries are submitted to MEDLINE. It is possible to overload the MEDLINE/Entrez web site if your site has a fast internet connection. To prevent this, the first 500 queries to MEDLINE are launched at full speed. After the first 500 queries, queries are launched no faster than one per second.

• New Features in Version 2.0

Improved Search Algorithm

The search algorithm has been changed substantially in version 2 of PDQ_MED. In the old version, MEDLINE was searched exhaustively with all possible pair-wise combinations of the query terms. In version 2, PDQ_MED downloads lists of all the abstracts that contain individual query terms. These lists are then cross-referenced to generate the raw co-occurrence matrix. If proximity searching is not being done, then PDQ_MED has all the information needed to complete the analysis. In the old version $\sim N^2/2$ searches were required. In this version, if proximity checking is not being done, only N searches are required.

If the user has requested a proximity search then this raw co-occurrence data is used to direct the researching of MEDLINE. Since the co-occurrence matrix is usually fairly sparse (contains a lot of zeros) significantly fewer searches are performed in version 2 of PDQ_MED.

Typically the new algorithm provides anywhere from a two to thirty fold speed increase compared to the old algorithm.

Indexing Limit

Since the new search algorithm retrieves a list of PMIDs for each query, a new parameter has been added to control the length of this list. The Maximum Abstracts to Index parameter limits the number of PMIDs returned for each query.

Suppressing the Graph Display

The user now has the option to suppress the graph display. The graph can be obtained by using the Resubmit Graph button from the output page. For more information, see Display Graph.

"Review" Links Removed

Since review articles containing multiple query terms are relatively rare, we have dropped the URLs that search for them specifically.

New MEDLINE URLs

We have updated the URLs that **PDQ_MED** uses to search MEDLINE, as well as the links to MEDLINE in the output pages, to the new PUBMED format.

Improvements in the Graph Display

The graph algorithm used has been changed and now provides a more stable graphic display. In addition, we have added two new options. The first option toggles the display between color and black and white. The second option allows the user to start and stop the minimizer.

• Quick Start Instructions:

PDQ_MED searches the more than 10 million records in PUBMED using the terms and phrases you supply. In the main text box, enter your terms such that each line represents a single gene, protein or concept. For genes or proteins with more than one name, enter the names separated by spaces on the same line. Terms with imbedded spaces need to be enclosed in quote marks. Terms on the same line are OR'd together, individual lines are pairwise AND'd together.

For example;

```
HGH "human growth hormone"
estradiol estrogen oestrogen 17b-estradiol
Interleukin-1b Interleukin-1beta IL-1b
```

Start the run by clicking on the "Submit" button at the bottom of the page.

Throughout the input form the  icon links to the relevant section in the online help manual.

• Input

Query format:

The query may be entered in the main text box or uploaded from a local file. In both cases, the format is the same.

Protect spaces in phrases with quote marks, e.g. "interleukin 1", otherwise words and phrases separated by spaces will be OR'd together. You can also can use AND, OR, or NOT but the spaces must be converted to +'s;

```
zag+NOT+zig
```

Enter the data so that each line contains the name and pseudonyms for a single gene, protein or concept. For example, to search for linkages between estrogens, estrogen receptor and the disease osteoporosis, a user might enter the following term list.

```
estrogen estradiol oestrogen 17b-estradiol "17beta estradiol"
"estrogen receptor" ER
osteoporosis osteopenia OP "bone disease" "hip fracture"
```

It is important to remember that pseudonyms and acronyms are not necessarily unique to a single gene or protein. Frequently acronyms are common English words, such as the acronym "ZAG" for zinc alpha-2 glycoprotein, which can lead to irrelevant linkages.

Term Tags:

Term tags allow you to mark terms with information about the query. Tags can be placed anywhere in a query line. The tags are enclosed in double square brackets [[ ]]. Multiple tags should be separated by semicolons and enclosed in a single set of square brackets. For example [[similar to; up;]]. Sample tagged queries are;

```
[[disease]] "Acute Myeloid Leukemia" "Acute Myelogenous Leukemia"
apoptosis[[state]]
[[up]] CD33
Zyxin[[down]]
[[enr]]"leptin receptor"
[[similar]]MAD2
```

The supported tags and how they affect your PDQ_MED run are shown in the table below.

Tag	Description	Effect on PDQ_MED	Graph Display
[[disease]]	disease or other non-gene query	State and disease terms are treated the same. The user may select to skip disease-to-disease, disease-to-state and state-to-state searches. In addition, PDQ_MED creates a separate output file for each disease or state term containing the linkages found to the particular state or disease term.	gray oval
[[state]]	A state (e.g., apoptosis, proliferation ...) or other non-gene query		
[[up]]	up regulated gene		green pentagon with a flat bottom and an upwards pointing vertex
[[down]]	down regulated gene		blue pentagon with a flat top and a downward pointing vertex
[[ENR]]	"Expressed but Not Regulated" (ENR) gene	The user may select to not do ENR-to-ENR searches.	gray rectangle
[[similar to]]	gene is similar to the named gene	Reminds the user that a proper name was not available for the gene.	pink rectangle
no tag	default		yellow rectangle

Query Comments:

Queries, in the paste box or loaded from a file, can contain comments delimited by "##". Everything on a line after the ## will be ignored by PDQ_MED. For example;

```
## Query list of 3 Dec. 2002
Fos
Jun
Lysozyme
catalase cat ## cat is not a very specific term for catalase
## ADAMTS1
```

The above query list is searched as if it was entered as;

```
Fos
Jun
Lysozyme
catalase cat
```

Local Pseudonyms and Aliases:

For **Sentence Proximity** and **Pharma Terms** checking a simple check for local pseudonyms/aliases is performed. Any term in parenthesis following a query term is regarded as an additional query term for that abstract. For instance, if a query for *raloxifene* returned an abstract containing *"... raloxifene (RAL) ..."* then, for that abstract only, the term *RAL* would also be used for the proximity checking.

Terms to avoid:

There are a few types of terms that should be avoided. The first is simple acronyms such as CAT (used for both chloramphenicol acetyltransferase and catalase) that will return abstracts relating to the proteins and, of course, felines. The second type of term to be careful with are the names of genes and proteins which are used as controls, markers or other parts of common experimental protocols such as luciferase, GapDH or polymerase. In addition, it is recommended that you avoid words that are common in the biomedical literature such as gene, cDNA, RNA, mRNA, cloning etc.

• Options

MEDLINE Options:

SEARCH FIELDS:

PDQ_MED supports many of the MEDLINE/PUBMED search fields. For a complete description of the PUBMED query system visit `http://www.ncbi.nlm.nih.gov/entrez/query/static/help/pmhelp.html`. The supported search fields are;

PUBMED Field Name	Description	Notes
ALL	All	
WORD	Text Word	
TITL	Title Word	
TIAB	Title/Abstract Word	**default**
SUBS	Substance Name	
MAJR	MeSH Major Topic	see Mesh Browser at www.ncbi.nlm.nih.gov /entrez/meshbrowser.cgi
MESH	MeSH Terms	
AUTH	Author Name	
ECNO	EC/RN Number	
PDAT	Entrez Date	
DP	Date of Publication	2001/01/01:2001/04/01[DP]
JOUR	Journal Name	
PDAT	Publication Type	
VOL	Volume	

The Search Field option is global and applies to all of the query terms. If you would like to specify a particular search field for one of your query terms use for example;

```
Jones[AUTH]
estrogen[SUBS]
Nature[JOUR]
CD33+AND+2001/01/01:2001/04/01[dp]
  (publication dates between Jan. 1 and March 1, 2001)
```

Note: Using search fields for individual terms will not work when doing proximity or pharma searches.

PDQ_MED Options:

LOCAL DATABASE OPTION

If your site has a local copy of the MEDLINE database then you have the option to use the local database instead of the NLM/NIH MEDLINE (Entrez) database. The various options to enable the local database are defined in the site.data file (contact your system administrator). If the local database is enabled then the following set of options are present.

Use Local Database:	Θ yes O no	localDB_Type:	mysql
		localDB_Address:	localhost
		localDB_Database:	localMEDLINE
		localDB_Table:	MEDLINE_abs
		localDB_TitleF:	TITLE
		localDB_AbsF:	ABSTRACT
		localDB_PmidF:	PM_ID
		localDB_Username:	<username>
		localDB_Pass:	<password>

Select the "yes" check box to use the local database. The localDB_Type is the type of database that is being used, supported values are "mysql" and "Oracle". The localDB_Address is the systems path to the database and the localDB_Database is the name of the database. The localDB_Table is the name of the table in the database that contains the various MEDLINE records, in particular, the abstract title (localDB_TitleF), the abstract (localDB_AbsF) and the PMID (localDB_PmidF). If a username and password are required to access the local database, then the localDB_Username and localDB_Pass (password) fields are also present.

There are several differences between searching a local database compared to a web search against MEDLINE.

1. The local database searching will not include the MeSH mapping (expanding of certain names and phrases) that Entrez does automatically.
2. The local database will not use the phrase index that Entrez uses. This is an advantage in that the query "foo bar" in the local database will only return records containing the

phrase. Entrez will look in its phrase index, and not finding this particular phrase, will re-map the query to "foo AND bar".

REPORT FORMAT STYLE

PDQ_MED has two basic report styles, "Groups" and "Lists". The "Groups" report style organizes the results based on the groups that the greedy clustering identified. In general, the clustering algorithm will find one large group, consisting of 25-75% of the terms, along with one or two small groups and several singleton groups. The majority of the results will be for the one large group, which is always called "Group 1".

In the "Lists" analysis the grouping of terms is based primarily on co-occurence with a "master" disease or state term. This analysis is meant to highlight connections between your list of genes and a particular disease or biological state. The first term in your query list that was tagged as a disease or state term is automatically selected as the "master" by PDQ_MED. The search results are then used to generate four lists of terms.

List-1 includes those genes that co-occur with the master disease/state term. That is, the gene co-occurs in sentences with the master term. This list summarizes what has been previously published concerning these genes and the master disease or state. It is expected that at least some of the genes in a typical microarray experiment should have literature precedence for their involvement in the disease or process. This list should provide a summary of what is already known concerning these genes and this biological process.

List-2 contains genes that are indirectly linked to the master term. That is, the gene is linked (co-occurs) with another gene, which in turn is linked to the master term. For the genes on this list, precedence for their involvement in the biological process was not found in MEDLINE. However, for these genes it may be possible to construct a rational explanation for their involvement in the particular biological process based on the terms with which they co-occur which in turn co-occur with the master term.

List-3 contains the genes that could not be linked, directly or indirectly, with the master term but have significant bodies of literature (five or more abstracts) or that co-occur with another gene that is not a member of List-1 or List-2. The genes on this list represent apparently novel findings concerning these genes and this biological process.

List-4 contains those genes that could not be linked, directly or indirectly, with the master term and which have little or no literature (less than 5 abstracts). The genes on this list represent the greatest challenge in rationalizing their involvement in the disease or process. Since there were no co-occurrences with other genes identified in the experiment (or the biological process) and there is little literature describing the gene, it is likely that little is known about the normal function of these genes.

If you are particularly interested in the literature relationships between your list of genes and a particular disease or biological process, then the "Lists" report is probably preferred over the "Groups" report.

GLOBAL TERM

The "Global Term" is an additional term that will be added to all of the searches. Pairs of terms are searched in MEDLINE along with the global term as;

```
Term-1 AND Term-2 AND GlobalTerm
```

The Global Term can be used to restrict the searches to only those abstracts that refer to a particular domain. For example, typical global terms might be;

```
cancer
CNS "central nervous system" neuron neuronal
```

The global term list is converted to an OR'd list in the same way as regular queries. In the example above, the CNS global term is searched as;

```
CNS OR "central nervous system" OR neuron OR neuronal
```

ONLY DO PROXIMITY SEARCHES FOR THE FIRST N TERMS

The user has the option to only do pair-wise searches for the first N terms in the query list vs. all of the remaining terms. Searches of pairs of terms where both terms are beyond the first N terms in the list are skipped.

This option is useful for searching a few terms against a large list of terms where the user is only interested in co-occurrences that contain a member from the short list and a member of the long list. If the user is particularly interested in only a few genes and would like to find co-occurrences with members of a large list of genes, perhaps all the highly expressed genes in a particular experiment, then this option is helpful. Consider the following query list;

```
FOS cFos hFos c-Fos AP1 AP-1 "ONCOGENE FOS"
NGFIA Egr1 G0S30 NGFI-A TIS8 G0S-30 KROX24 ZIF-268 EGR-1 TIS-8
A-BETA
ABL1 ABL ABL-1 JTK-7
ABL2 ABLL ARG "ABELSON-RELATED GENE" ABL-2
ACAC ACC ACACA ACACA ACCA ACC1
acetylcholinesterase "acetylcholinesterase acetylcholine" ACHE
ACHRA CHRNA-1 CHRNA "acetylcholine receptor"
"ACTIN BETA" ACTB "BETA ACTIN" BETA-ACTIN
"ACTIVATING TRANSCRIPTION FACTOR 2" CRE-BP1 TREB-7 ATF2 TREB7
adaptin
...
XLP XLPD IMD-5 EBVS MTCP1 SH2D1A "SLAM-ASSOCIATED PROTEIN" SAP
YY1
ZFP
ZNF-161 DB-1 DB1
```

If the user sets the "Only do Proximity searches for the first N terms:" option to 2 then the first two terms, "FOS.." and "NGFIA...", will be searched against all other terms in the query list. (They will also be searched against each other.) However, all of the pairwise searches that do not include one of the first two terms will be skipped.

STATE TO STATE SEARCH

The "State to State Search" option instructs PDQ_MED to not search for co-occurrences between terms that were both tagged as disease/state terms in the query list. This option is useful for finding relationships between a set of genes and a set of disease or state terms. If you enter a set of genes and a list of disease names (each tagged as [[disease]]) this option will search the genes vs. the disease but skip the disease vs. disease searches. For example;

```
#########################
## Genes
fumarylacetoacetase "fumarylacetoacetate hydrolase"
zyxin
"LTC4 synthase" "LTC4 synthetase"
LYN
HoxA9 "Hox A9" "Hox-A9"
cd33
adipsin ADN "Complement Factor D" "D component of complement"
```

```
"leptin receptor"
########################
## "States" and "Diseases"
[[state;]] "cell cycle" mitosis
  [[state;]] G0
  [[state;]] G1
  [[state;]] G2
  [[state;]] "S phase"
  [[state;]] "M phase" metaphase
[[state;]] apoptosis apoptotic "programed cell death" "programmed cell death"
[[state;]] cancer
  [[state;]] "breast cancer"
  [[state;]] "ovarian cancer"
  [[state;]] "colon cancer"
  [[state;]] "prostate cancer"
  [[state;]] "skin cancer" melanoma
  [[state;]] "lung cancer"
  [[state;]] leukemia leukemic
  [[state;]] metastasis metastasize metastasized metastatic metastases
[[state;]] cardiovascular heart "smooth muscle"
[[state;]] toxicity toxic liver toxicities
[[state;]] alzheimer parkinson amyloid
[[state;]] bone skeleton skeletal bones
  [[state;]] osteoblast osteoclast osteocyte
  [[state;]] osteoporosis osteopenia
  [[state;]] pagets paget osteopetrosis paget's
[[state;]] immune immuno
  [[state;]] complement antibody
  [[state;]] infection
  [[state;]] inflammation swelling
  [[state;]] T-cell T-cells "T cell" "T cells" macrophage lymphocyte lymphoid
  [[state;]] hiv aids
  [[state;]] SARS
```

This query will identify linkages between the eight genes at the top of list and any of the 29 state/disease terms. The searches between the 29 state/disease terms will be skipped. This type of search may help to classify a particular set of genes into a disease or biological process category.

ENR TO ENR SEARCH

The "ENR to ENR Search" option instructs PDQ_MED to not search for co-occurrences between terms that were both tagged as [[ENR]] (Expressed but Not Regulated) in the query list. This allows the user to enter the up and down regulated genes as well as the highly expressed genes ([[ENR]]) as queries. Since there are typically many more highly expressed genes than up/down regulated genes, a regular search will tend to be dominated by searches between the ENR genes. This option allows the user to extract value from the ENR genes, by finding co-occurrences with the regulated genes, without being swamped by the ENR-ENR co-occurrences.

PROXIMITY SEARCH

Proximity searching takes the basic PDQ_MED algorithm one step further. When proximity searching is turned on every abstract that contains two of the query terms is downloaded and examined to see if the two terms occur in the same sentence, instead of just in the same abstract. Proximity tends to find much closer association between two terms than does the default search. Proximity takes considerably longer to run since all of the abstract must be downloaded to the PDQ_MED server, which can be time consuming. Proximity searching

uses any local pseudonyms and acronyms found in a particular abstract (see the section on Local Pseudonyms and Aliases).

PHARMA TERMS SEARCH

Pharma Terms search is similar to proximity search. However, instead of requiring two of the query terms in a single sentence, Pharma terms highlights sentences that contain a query term and a term from the Pharma list. The Pharma list is designed to highlight the most critical types of relationships, ones that are of greatest potential value to a pharmaceutical researcher.

If desired, you can alter the list of pharma terms in the text box. Note that the Pharma terms format is somewhat different than for the query terms. For Pharma terms, each term is considered unique (unless enclosed in quote marks) and the entire set is converted to an OR list and searched as one term.

Since the pharma search is a kind of proximity search, it takes longer to run than a basic search.

MAXIMUM ABSTRACTS TO INDEX

Prior to generating the cross-reference list, PDQ_MED downloads the list of PubMed ID's (PMID) for each query phrase. This parameter limits the number of ID's that will be returned. Increasing this parameter will increase PDQ_MED sensitivity at the expense of speed. Decreasing this parameter will increase speed at the cost of sensitivity.

The default value of 20,000 ID's is a reasonable trade-off, favoring sensitivity over speed.

MAXIMUM ABSTRACTS TO CHECK

NOTE: This parameter only affects Proximity and Pharma searches.

If you are doing a Proximity or Pharma term search, then the Maximum Abstracts value controls how many abstracts will actually be downloaded for analysis. The default value of 50 usually provides a good trade off between speed and sensitivity. The maximum value, 5000 abstracts, is a limitation imposed by MEDLINE / PUBMED.

GROUPING TYPE & GROUPING CUTOFF

After all possible pair-wise term searches have been completed, PDQ_MED uses a greedy clustering algorithm to group the terms. If term-A and term-B co-occur in a set of abstracts and term-B and term-C co-occur in another set of abstracts, then term-A, term-B and term-C are all members of the same group. The Grouping Type and Grouping Cutoff options allow the user to control this process.

The Grouping Type option has two settings; Raw and Weighted. If the Grouping Type is set to Raw then the measure of the co-occurrence between two terms is simply the number of abstracts in which they co-occur. If Grouping Type is set to Weighted then the measure of the co-occurrence between two terms is defined as;

$$\frac{(Term-A)\bigcap(Term-B)}{\text{minimum}(Term-A, Term-B)}$$

For example, if Term-A occurs in 300 abstracts, Term-B occurs in 5 abstracts and the two terms co-occur in 2 abstracts then the Weighted measure of the co-occurrence is;

$$\frac{(Term-A)\bigcap(Term-B)}{\text{minimum}(Term-A, Term-B)} = \frac{2}{\text{minimum}(300,5)} = \frac{2}{5} = 0.4$$

Using a Weighted measure changes the relative importance of the total number of co-occurrences vs. the fraction of the total number of abstracts that contain either term. In the above example, even though there are only two abstracts that contain both terms, it represents 40% of the abstracts that contain the rarer term (Term-B) which suggest a significant linkage.

The Grouping Cutoff value controls the minimum number (or fraction) of co-occurrences that are required for grouping. If the Grouping Type is "Raw" then the default value is 1, which means all co-occurrences are significant. If you increase this value then you require a larger number of co-occurrences in order to trigger grouping. If Grouping Type is "Weighted" then the Grouping Cutoff represents the minimum fractional co-occurrence for grouping.

DISPLAY GRAPH

This option allows the user to suppress (turn off) the Graph display in the output. If you are doing a very large (more than 100 queries) unattended run, you may want to suppress the graph display to insure that the browser will continue to function properly.

If you suppress the graph display, you can still display it from the output page using the Resubmit Graph button.

The default is to display the graph.

• Description of the Output

Run Statistics

RUN TITLE

The first segment of the output gives the PDQ_MED version number, run title, date and time. The "011109124218auser13221" translates as 2001/11/9 12:42:18 PM (time at the server) username and process ID.

```
Project type: PDQ_MED
Version: 2.81, 23 Sep 2003
Host Site: InPharmix Inc. This is a full license for InPharmix Inc. software which expires 31 Dec 2001.
Data set: 031009124218auser13221
```

RUN PARAMETERS

The run parameters block displays the various user selected options, statistics about the number of MEDLINE searches that will be done and gives an estimate of the expected run time.

```
                Total query phrases = 9
                Report Format Style = Groups and Lists
                        Master Term = "Acute Myeloid Leukemia" OR "Acute Myelogenous!
Maximum number of indexes per query = 20000
 Maximum potential MEDLINE searches = 45
                      MEDLINE field = tiab
               State to State search = yes ("State" includes state and disease tags)
                   ENR to ENR search = no  (ENR = "expressed but not regulated" tag)
                    Proximity search = yes
                       Pharma search = yes
         Maximum abstracts to check = 250

                          GlobalTerm = no global term
          Limit Pair Wise Searches = no limit

                       Grouping type = raw
                     Grouping cutoff = 1

                 Estimated run time =  6.9 minutes.
```

RUNNING LOG

As the PMID list for each term is retrieved from MEDLINE, the results are presented in this table.

The columns, from left to right, give;

1. The search number
2. The number of hits (PMIDs returned) for a single query phrase which is also a link to the MEDLINE results
3. The query phrase

```
Indexing query terms (index limit = 20000 per term) ...
-------------------------------------------------------------------------------
Search   Hit
   #    Count     Term
-------------------------------------------------------------------------------
   1     8942     "Acute Myeloid Leukemia" OR "Acute Myelogenous Leukemia"
   2       66     HoxA9 OR "Hox A9"
   3     1315     CD33
   4      819     "leptin receptor"
   5       97     Zyxin
   6       51     MAD3
   7      162    [[similar to]] MAD2
   8      144     Adipsin
   9        0     Nohitsforthisterm
-------------------------------------------------------------------------------
Done.  (index time = 2 seconds)
Total abstracts in MEDLINE for these terms = 11596
               Average abstracts per term = 1288.4
         Expected number of co-occurrences =    0.3  (null hypothesis)
```

RUNNING PAIRWISE LOG

As PDQ_MED cross-references each term pair, the results are shown in this table. If a proximity search is being done, then the results of the additional MEDLINE searches is also shown.

The columns, from left to right, give;

1. The search number
2. The percent of the total searches which have been completed

3. The number of hits for a single query or the number of co-occurrences for paired queries
4. The search term which is also a link to the MEDLINE results
5. The proximity search results (if done). For example "99 of 250 (278)" means there was a total of 278 abstracts but the "Maximum Abstracts to Check" setting limited the actual checking to the first 250 abstracts of which 99 had the terms in the same sentence.

```
Pairwise searching MEDLINE (5 searches to do) ...
-------------------------------------------------------------------------------
  Search    %
     #   Complete  Hits    Terms                                    Proximity
-------------------------------------------------------------------------------
     0      0%     8942    "Acute Myeloid Leukemia" OR "Acute Myelogenous Leukemia"
     1     20%       15       AND HoxA9 OR "Hox A9"                 15 of 24
     2     40%       99       AND CD33                              99 of 250 (278)
     3     60%        1       AND "leptin receptor"                  1 of 1
     3     60%       66    HoxA9 OR "Hox A9"
     3     60%     1315    CD33
     3     60%      819    "leptin receptor"
     4     80%        0       AND Adipsin                           0 of 1
     4     80%       97    Zyxin
     4     80%       51    MAD3
     5    100%        9       AND [[similar to]] MAD2               9 of 9
     5    100%      162    [[similar to]] MAD2
     5    100%      144    Adipsin
     5    100%        0    Nohitsforthisterm
-------------------------------------------------------------------------------
9 single term and 5 double term = 14 total searches.
Done.  (pair-wise search time = 9 seconds)
```

RUNNING PHARMA LOG

As each Pharma search is done, the results are presented in this table. The columns, from left to right, give;

1. The search number
2. The percent of the total searches which have been completed
3. The number of hits for the query and the Pharma Terms
4. The search terms. The AND links to the search results.
5. The proximity search results (if done). For example "149 of 250 (4480)" means that the returned abstracts were limited to 250 (there were a total of 4480) and of the examined 250 abstracts which contained both terms, 149 had the terms in the same sentence.

```
Doing the Pharma searches ...
-------------------------------------------------------------------------------
  Search    %
     #    Complete  Hits   Terms                                      Proximity
-------------------------------------------------------------------------------
     1      11%      149    "Acute Myeloid Leukem! AND  Pharma Terms  149 of 250 (4480)
     2      22%       18    HoxA9 OR "Hox A9"      AND  Pharma Terms   18 of 36
     3      33%      105    CD33                   AND  Pharma Terms  105 of 250 (659)
     4      44%      118    "leptin receptor"      AND  Pharma Terms  118 of 250 (601)
     5      56%       50    Zyxin                  AND  Pharma Terms   50 of 85
     6      67%       26    MAD3                   AND  Pharma Terms   26 of 43
     7      78%       67    MAD2                   AND  Pharma Terms   67 of 105
     8      89%       55    Adipsin                AND  Pharma Terms   55 of 107
skipping Nohitsforthisterm (no hits)
-------------------------------------------------------------------------------
Done.
```

"Groups" Output Format

Once the searches of MEDLINE have been completed, PDQ_MED clusters (groups) the terms based on their co-occurrences in MEDLINE. The next section of the output contains the "Groups" style report format. This section is only shown if the "Report Format Style" option was "Groups" or "Groups and Lists". A separate section of output is presented for each of the groups. The individual sections include the co-occurrence data tables, graph and proximity sentences for a single group.

MEMBERS SUMMARY

The term clusters table lists the number of members in each group. PDQ_MED will always assign the largest group as Group 1. Hyperlinks to the co-occurrence data files are also given.

```
Finding Term Clusters ... Done.
Group number 1 has 4 member terms.
Group number 2 has 2 member terms.
And, there are 3 single term groups.

Writing to output files ... Done.
Abstract level co-occurrence matrix written to this file.
Normalized co-occurrence matrix written to this file.
```

MEMBERS PER GROUP SUMMARY

The next part of the PDQ_MED output consists of several sections of data for each group.

In the table below, the data for the first group (1) is shown. Note the message reminding the user that only 50 abstract were checked for proximity.

Every query term in the group has its own sub-table and every term is listed as both the first and second term in a search.

The columns, from left to right, give;

1. The number of hits for this term (Num ABS)
2. The number of co-occurrences for this term pair (Co-occur Count)
3. The normalized co-occurrences for this term pair (Norm)
4. The term. Underlined terms link to MEDLINE. The AND links to MEDLINE for that term pair.
5. The proximity results (Proximity)

Group 1: 4 members.

Num ABS	Cooccur Count	Norm	Term	Proximity
Abstracts checked limited to 250 candidate abstract per term pair.				
8942			**"Acute Myeloid Leukemia" OR "Acute Myelogenous Leukemia"** (*StAT*)	
1315	99*	0.083717*	AND (*StAT*) CD33	99 of 250 (278)
66	15	0.227273	AND (*StAT*) HoxA9 OR "Hox A9"	15 of 24
819	1	0.001221	AND (*StAT*) "leptin receptor"	1 of 1
1315			**CD33** (*StAT*)	
8942	99*	0.083717*	AND (*StAT*) "Acute Myeloid Leukemia" OR "Acu!	99 of 250 (278)
66			**HoxA9 OR "Hox A9"** (*StAT*)	
8942	15	0.227273	AND (*StAT*) "Acute Myeloid Leukemia" OR "Acu!	15 of 24

```
   819   "leptin receptor" (StAT)
           1 weakly linked term(s) not shown (see here instead).
  8942     1  0.001221   AND (StAT)"Acute Myeloid Leukemia" OR "Acu!  1 of 1
--------------------------------------------------------------------------------
* Estimated because of "Maximum Abstracts to Check" limit.
```

Number of terms in group =	4
Total abstracts for terms in this group =	11,142
Number of term pairs =	3
Total number of sentences with two or more terms =	115

MINIMAL SPANNING TREE

The following section of the PDQ_MED output presents this group as a hierarchical tree. The base node of the tree is the pair of terms that have the highest co-occurrence measure, either the "raw" or "normalized" value depending on the setting of Grouping Type. Subsequent terms are added to the tree in order of the strength of co-occurrence with terms already present in the tree. Underlined terms (or term phrases) link to MEDLINE. The underlined number within the parenthesis is the -Log_{10} for the normalized co-occurrence measure for that term pair and links to MEDLINE. Smaller numbers indicate stronger links.

The tree analysis is useful for presenting a hierarchical view of the terms. However, this analysis has some important limitations. In particular, a given term can link **upwards** in the hierarchy to only one other term. For example, in the output below, even though *HoxA9* and *CD33* co-occur, they are not shown as being linked in the hierarchy since they both co-occur more frequently with *"Acute Myeloid Leukemia"* than they do with each other.

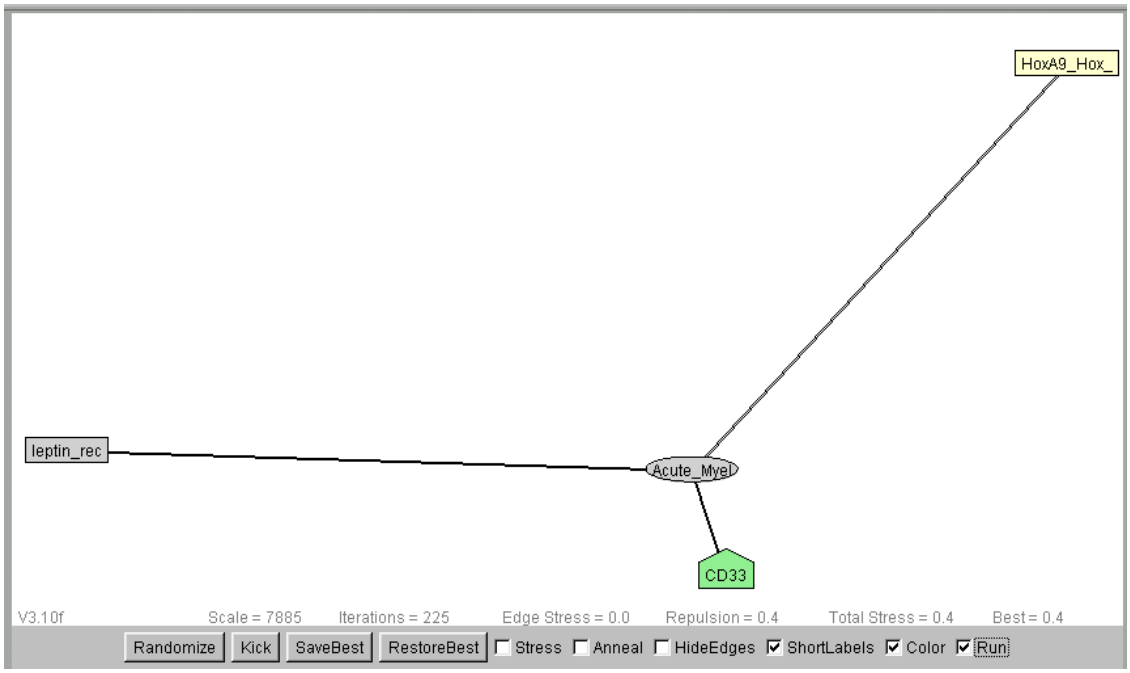
```
Best term pair = "Acute Myeloid Leukemia" OR "Acute Myelogenous Leukemia"
   AND HoxA9 OR "Hox A9".
Matching at 15 (raw) and 0.227273 (normalized).


"Acute Myeloid L!---(0.6)HoxA9 OR "Hox A9!
   |
   |----------------(1.1)CD33
   |
   |----------------(2.9)"leptin receptor!
```

GRAPH

An alternative to the Tree presentation is a distance geometry treatment of the data. In this graphical analysis of the data, the inverse of the co-occurrence data is used as distance constraints and a solution to this set of constraints in two-dimensional space is sought. In the graph display, each term (or term phrase) is presented as a box. Terms that co-occur are linked by a line, the target length of which is proportional to the inverse of the co-occurrence frequency (frequent co-occurrences have short lines, infrequent co-occurrences have longer lines). Satisfied edges are shown as black lines, unsatisfied edges as red lines.

The Java applet automatically centers and scales the displayed image. In addition, there are several controls for the applet to help the user explore their data.

1. Boxes can be repositioned by dragging with the left mouse button.
2. Boxes can be fixed in place with the left + right mouse buttons.
3. The shape and color of the boxes is determined by the tags (disease, up regulated etc.) present in the input query list.
4. *Scale:* The current scaling factor for the display with 1000 indicating 100%. This line of parameters is only shown when the Java applet is actually running.
5. *Iterations:* The number of iteration (optimization) cycles that have been completed. By default, the optimization stops after 1001 iterations. The "Run" checkbox can be used to continue the optimization.
6. *Edge Stress Value:* Displays the total "energy" contribution from unsatisfied distance (edge) constraints.
7. *Repulsion:* The "energy" contribution from "repulsion", the portion of the energy function which tries to spread the nodes out as much as possible.
8. *Total Stress Value:* Displays the total energy (edge + repulsion) of the current configuration. This is the value the graph algorithm is trying to minimize.
9. *Best:* The best energy value encountered so far. The ResetBest button resets this value.
10. *Randomize Button:* Randomly distributes the nodes in the display window. This is the configuration at the beginning of a minimization run.
11. *Kick Button:* Randomly displaces all nodes by a moderate amount. This button is useful for trying to "untangle" a complex graph.
12. *SaveBest Button:* Saves the current configuration as the "Best" configuration.
13. *RestoreBest Button:* Restores the best configuration found since the last SaveBest.
14. *Stress Checkbox:* Display the stresses associated with each edge. Displays the values associated with each edge. For example, a display along an edge of "1.0;d=1.3; 6.0; d=0.07" indicates the terms co-occurred 1 time, the length of the currently displayed edge corresponds to a co-occurrence of 1.3 times, in the internal units of the graph (which ranges from 1 to 6) the edge is of length 6 and its error is 0.07 (out of 6).
15. *Anneal Checkbox:* Initiates a simple "simulated annealing" algorithm to try to locate the optimal configuration for complex graphs. The configuration is periodical "kicked" and then optimized. If a new best configuration is found it is saved. The best configuration encountered can be restored with the *RestoreBest* button.
16. *HideEdges Checkbox:* Hides the display of the edges (lines).
17. *ShortLabels Checkbox:* Toggles between the display of the full term names in the boxes vs. only the first ten characters of each name.
18. *Color Checkbox:* Toggles the display between color and black and white.
19. *Run Checkbox:* Starts and stops the minimizer.



REGRAPH

The ReGraph button submits the data for this group to the ReGraph processor. The ReGraph processor allows the user to transform, filter and further explore the relationships within this group. See the section on GraphInPharmix latter in this document.

KEY SENTENCES

The Key Sentences display lists sentences that contain three or more query terms or two or more query terms plus a Pharma term. The underlined blue number is a MEDLINE link to the full abstract. Query terms are in **bold face** and Pharma terms in green. (Only shown if a Proximity Search was done.)

"Key" Sentences for Group 1:

("Acute Myeloid Leukemia" OR "Acute Myel!) ***AND*** *(CD33)* ***AND*** *(PharmaTerm)*

11986941 Gemtuzumab ozogamicin (CMA-676), a calicheamicin-conjugated humanized anti-CD33 mouse monoclonal antibody, has recently been introduced clinically as a promising **drug** for the treatment of patients with **acute myeloid leukemia** (AML), more than 90% of which express **CD33** antigen.

12042706 Anti-**CD33** antibodies have been used alone-and more effectively, attached to chemo**therapy** agents or radioisotopes-to treat those with **acute myeloid leukemia**.

12200674 We analyzed the safety and **efficacy** of Mylotarg (gemtuzumab ozogamicin, an antibody-targeted chemotherapy consisting of a humanized anti-**CD33** antibody linked to calicheamicin, a potent antitumor antibiotic) in the treatment of 101 patients > or =60 years of age with **acute myeloid leukemia** (AML) in untreated first relapse in three open-label trials.

12447847 Gemtuzumab ozogamicin, a calicheamicin-conjugated anti-**CD33** monoclonal antibody, has demonstrated substantial **efficacy** in patients with **acute myeloid leukemia** (AML) and has induced remissions in patients with favorable-, intermediate-, and poor-risk cytogenetics.

12689934 To address whether multidrug resistance protein (MRP) affects GO susceptibility, we characterized Pgp, MRP1, and MRP2 expression in CD33+ cell lines and **CD33**+ **AML** samples and analyzed the effect of the Pgp inhibitor cyclosporine (CSA) and the MRP **inhibitor** MK-571 on GO-induced cytotoxicity.

12700948 Gemtuzumab ozogamicin, a calicheamicin conjugate that **targets** **CD33**, has recently been approved by the Food and Drug Administration (FDA) for treatment of **acute myelogenous leukemia** (AML).

PROXIMITY SENTENCES

The Proximity Sentences display list the sentences that contain two query terms. The underlined blue number is a MEDLINE link to the full abstract. The number in parenthesis following the MEDLINE link is the ***StAT*** score, which is used to rank the sentences in decreasing order of "importance". For further information, see the ***StAT*** portion of this manual. Query terms are in shown in **bold face**. (Only shown if a Proximity Search was done.)

"Acute Myeloid Leukemia" OR "! **AND** *CD33*

12592328 (22.8) TITLE: Treatment of relapsed or refractory **acute myeloid leukemia** with humanized anti-**CD33** monoclonal antibody HuM195.
12141950 (20.2) TITLE: Phase III trial of a humanized anti-**CD33** antibody (HuM195) in patients with relapsed or refractory **acute myeloid leukemia**.
13680159 (19.1) Mylotarg, a humanized anti-**CD33** antibody linked to an antitumor antibiotic, is approved for the treatment of patients with relapsed **acute myeloid leukemia** (AML).
12451477 (17.9) PURPOSE: Mylotarg, a humanized anti-**CD33** antibody linked to an antitumor antibiotic, is approved for the treatment of patients with relapsed **acute myeloid leukemia** (AML).

12700948 (12.1) Gemtuzumab ozogamicin, a calicheamicin conjugate that targets **CD33**, has recently been approved by the Food and Drug Administration (FDA) for treatment of **acute myelogenous leukemia** (AML).
11474494 (11.9) The antigen **CD33** is expressed on blast cells in 80% to 90% of **acute myeloid leukemia** (AML) cases but, importantly, is not expressed on pluripotent hematopoietic stem cells or on nonhematologic cells.
12004080 (11.6) Anti-**CD33** antibodies are being used to deliver cytotoxic agents, such as calicheamicin to patients with **acute myeloid leukemia** with response rates up to 30%.

11342449 (8.8) TITLE: Targeting of the **CD33**-calicheamicin immunoconjugate Mylotarg (CMA-676) in **acute myeloid leukemia**: in vivo and in vitro saturation and internalization by leukemic and normal myeloid cells.
12447847 (8.8) Gemtuzumab ozogamicin, a calicheamicin-conjugated anti-**CD33** monoclonal antibody, has demonstrated substantial efficacy in patients with **acute myeloid leukemia** (AML) and has induced remissions in patients with favorable-,
.
.
.

Only the first 50 proximity sentences are shown for this term pair. There were an additional 49 sentences that are not shown.

At this point, all of the output for a single group has been presented. If there are additional groups then the complete set of output tables is repeated for each group.

PHARMA SENTENCES

If a pharma search was done then those results are displayed next. The Pharma Sentences display list the sentences that contain a query term plus a Pharma term. The underlined blue number is a MEDLINE link to the full abstract. Query terms are in **bold face** and Pharma terms in green. (Only shown if a Pharma Search was done.)

Pharma Sentences:

"Acute Myeloid Leukemia" OR "Acute Myelogenous Leukemia"

12883529 Autologous bone marrow transplantation is an alternative **therapeutic** option for **acute myeloid leukemia** patients lacking a compatible donor.
12883747 TITLE: Gleditsia sinensis fruit extract is a potential chemo**therapeutic** agent in chronic and **acute myelogenous leukemia**.
12884821 TITLE: [High dose ara-C **therapy** induced bradycardia in an **acute myeloid leukemia** patient with inv (16)(p13q22)].
12885813 PURPOSE: To evaluate the response rate, survival, and toxicity of mitoxantrone and cytarabine induction, high-dose cytarabine and etoposide intensification, and further consolidation/maintenance **therapies**, including bone marrow transplantation, in children with relapsed, refractory, or secondary **acute myeloid leukemia** (AML).

12921949 Vincristine (VCR) is an effective **drug** against acute lymphoblastic leukemia (ALL), many solid tumors, but not **acute myeloid leukemia**.
12923866 Probes have also been designed to hybridize to multiple cis paralogs, both enhancing the chromosomal **target** size and detecting chromosome rearrangements, for example, by splitting and separating a family of related sequences flanking an inversion breakpoint on chromosome 16 in **acute myelogenous leukemia**.
12928754 Over a 10-year period, 66 primary resistant **AML** patients have been treated by EMA salvage chemo**therapy**.
12930314 Therefore, timely identification and **therapeutic** stratification of those patients deemed at high risk for disease relapse could ultimately result in a further improvement of clinical outcome within these cytogenetic subgroups of **AML**.

12951753 Gemtuzumab ozogamicin is an active second-line **therapy** in older patients with **acute myelogenous leukemia**, but its role in combination regimens is unclear.
12956443 Previously, t(11;16) has been reported in fewer than 20 patients, all with the diagnosis of **therapy**-related (secondary) **acute myelogenous leukemia** (sAML) or myelodysplastic syndrome (MDS).
12962366 TITLE: The clinical relevance of the expression of several multi**drug**-resistant-related genes in patients with primary **acute myeloid leukemia**.
12963850 This finding has potential utility for **therapy** of patients with **AML**.
12966906 Gemtuzumab ozogamicin is an active second-line **therapy** in older patients with **acute myelogenous leukemia**, but its role in combination regimens is unclear.
.
.
.

Only the first 50 pharma sentences are shown for this term There was an additional 100 sentences that are not shown.

"List" Output Format

In the four lists analysis the grouping of terms is based primarily on co-occurence with a "master" disease or state term. This analysis is meant to highlight connections between your list of genes and a particular disease or biological state. The first term in your query list that was tagged as a disease or state term is automatically selected as the "master" by PDQ_MED.

List 1 includes those genes that co-occur with the master disease/state term. That is, the gene co-occurs in sentences with the state term. This list summarizes what has been previously published concerning these genes and the master disease or state. It is expected that at least some of the genes in a typical microarray experiment should have literature precedence for their involvement in the disease or process. This list should provide a summary of what is already known concerning these genes and this biological process.

List 2 contains genes that are indirectly linked to the master term. That is, the gene is linked (co-occurs) with another gene, which in turn is linked to the master term. For the genes on this list, precedence (based on sentence level co-occurence) for their involvement in the biological process was not found in MEDLINE. However, for these genes it may be possible to construct a rational explanation for their involvement in the particular biological process based on the terms with which they co-occur which in turn co-occur with the master term.

List 3 contains the genes that could not be linked, directly or indirectly, with the master term but have significant bodies of literature (five or more abstracts) or that co-occur with another gene that is not a member of List 1 or List 2. The genes on this list represent apparently novel findings concerning these genes and this biological process.

List 4 contains those genes that could not be linked, directly or indirectly, with the master term and which have little or no literature (less than 5 abstracts). The genes on this list represent the greatest challenge in rationalizing their involvement in the disease or process. Since there were no co-occurrences with other genes identified in the experiment (or the biological process) and there is little literature describing the gene, it is likely that little is known about the normal function of these genes.

INDEX

The Lists report Index table outlines the sections that will follow. The number of terms on each list is also given in the index.

COMPLETE CO-OCCURRENCE GRAPH

This graph shows all co-occurrences found for your set of queries. This graph is the same as the "Group 1" graph for the Groups output report format.

MEMBERS OF LIST N

The report now proceeds through sections of output for each of the four lists. For each list, the first output table is a list of the members of the particular list.

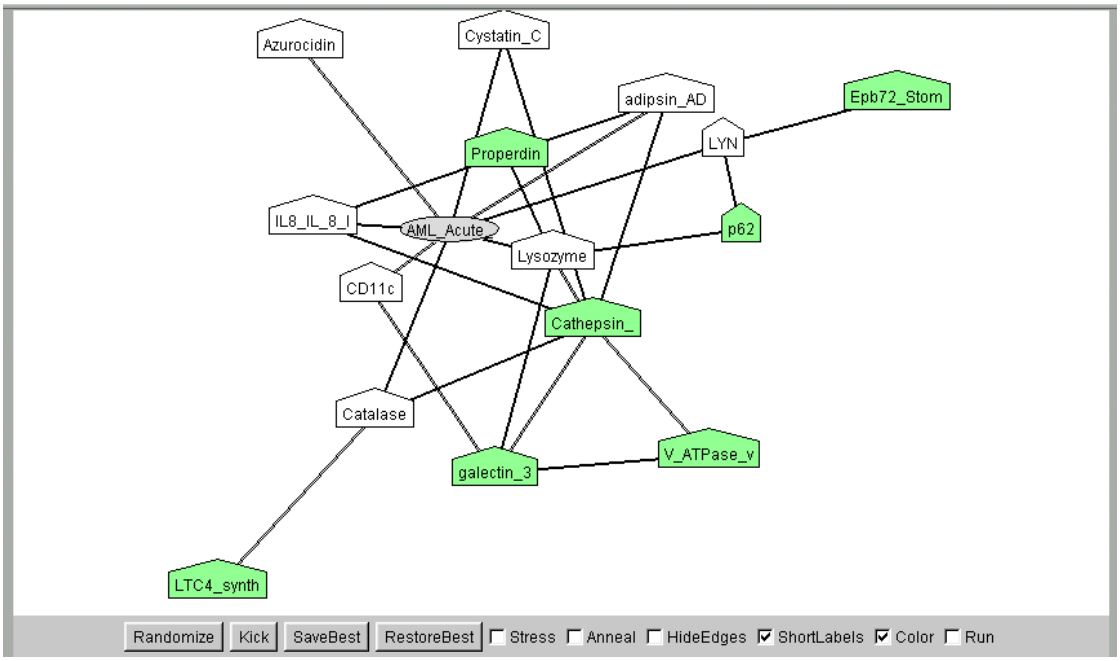
CO-OCCURRENCES OF MEMBERS OF LIST N

The co-occurrence statistics for members of this list are shown. The output is the same as described previously for the Groups output format.

The report for List 2 is somewhat different than for the other lists. Recall that List 2 members are not linked directly to the master disease/state term but instead are linked indirectly via another term. The co-occurrence data therefore is broken down into two subsets, co-occurrence between members of List 2 and co-occurrences between members of List 2 and List 1. The List 2 -- List 1 co-occurrences provide the path from the List 2 terms to the master disease/state term.

GRAPHINPHARMIX FOR LIST N

The graphs for List 1 and List 2 contain only the members of those lists. In addition, the List 2 output section has an additional graph that shows the linkages between the List 2 members and List 1. List 1 member nodes are uncolored and the connections between members of List 1 are not shown.



List 3 may have multiple graphs depending on the extent the terms co-occur. List 4 does not have a graph since it contains only unlinked terms.

PROXIMITY SENTENCES FOR LIST N

Each list, when appropriate, will have various sections showing the proximity sentences. The first section for a given list will give the "Key" sentences for that list. "Key" sentences include three or more query terms, or, two or more query terms plus a pharma term (when a pharma search is done). The next section gives all the two term proximity sentences for the particular list. The underlined blue number is a MEDLINE link to the full abstract. The number in parenthesis following the MEDLINE link is the StAT score, which was used to rank the sentences in decreasing order of "importance". Query terms are in shown in **bold face**.

Terms are in alphabetical order and each term pair is only shown once. To locate the proximity sentences between a pair of terms, first find the section for the term that occurs first alphabetically, and then look within that section for the second term. The sorting is done ignoring all non-letters and non-numbers (so quote marks don't affect sorting). The query pair "Aaaa" and "Cccc" will be shown in the Aaaa section only, and will not be shown in the Cccc section.

Concluding Analyses and Tables

STATE AND DISEASE SUMMARY FILES

PDQ_MED creates summary files for all of the terms that were tagged as either "disease" or "state". The file for a particular disease/state term contains the co-occurrence data and sentences containing the particular disease/state term.

Note: Summary files are only kept on the server for 10 days; you may want to save as a local file.

State and Disease Summary Files:
[[disease;]] Acute_Myeloid_Leukemia__Acute_

INSIGNIFICANT CO-OCCURENCES

A file containing the "insignificant" co-occurrences is created by PDQ_MED. This table lists term pairs for which there were co-occurrences but the co-occurrence failed proximity checking or the frequency of co-occurrence was less than the Grouping Cutoff.

The columns, from left to right, give;

1. The reason the term pair is considered insignificant, either proximity or cutoff.
2. The group (or list) the first term belongs to.
3. The number of abstracts containing the first term.
4. The first term.
5. The group the second term belongs to.
6. The number of abstracts containing the second term.
7. The second term.
8. The raw count of abstracts that contained the terms together.
9. Links to MEDLINE and StAT that search this term pair.

"Insignificant" cross references found for the terms:

Grouping:
Proximity Search = yes
Type = raw
Cutoff = 1

	Term 1			Term 2			Term 1 + Term2	
By	Grp	Count	Term 1	Grp	Count	Term 2	Raw	Links
proximity	3	144	[[down]] Adipsin	1	819	[[enr]]"leptin receptor"	1	MEDLINE, *StAT*

NO CO-OCCURENCES FOR THE TERMS:

The table below lists all terms that did not co-occur with any other terms based on proximity checking (if done) and the Grouping Cutoff. If terms on this list did co-occur, but were removed by the proximity or Grouping Cutoff criteria, then they are marked with "This term has cross-references that were omitted because of proximity of cutoff settings". Terms so marked are listed in the "Insignificant Co-occurrences" file (see above).

Term	# of abstracts	Links	
Adipsin This term has cross-references that were omitted because of proximity or cutoff settings.	144	MEDLINE	*StAT*
Zyxin	97	MEDLINE	*StAT*

NO ABSTRACTS FOUND IN MEDLINE FOR THE TERMS

This table lists the terms for which no hits in MEDLINE were found.

No abstracts found in MEDLINE for the terms;
Nohitsforthisterm (Search)

LOCAL ALIASES

The Local Aliases table reports on any local aliases found. Local Aliases are only searched for when Proximity Search or Pharma Searches are being done. The table below lists the local aliases found for each of your query terms. These aliases were used for proximity checking only in the abstract where they were defined. The columns for this table, from left to right, contain;

1. The query term for which the alias was found
2. The alias
3. The link to the MEDLINE abstract

Local Aliases

Term	Alias	Link(s) to abstract(s)
Acute Myelogenous Leukemia	*AML*	9322892, 9717827, 9777893, 10229319, 10397741, 10602420, 11041016, 11042523, 11049021, 11197213,
	AML-M2	9637886
	T-AML	9766650
estradiol	*E2*	6768442, 10523013, 11456279, 11509969
raloxifene	*EVISTA*	11535963, 11589065, 11770189
	RLX	11451623, 11589267, 11846327

TERM FREQUENCIES

The term frequencies table gives a summary of the term frequencies as well as their frequencies of co-occurrence. The columns in the table show;

1. `Total Abs:` The total number of abstracts in MEDLINE that contain this term.
2. `Co-occur Terms:` The number of terms that co-occur with this term.
3. `Co-occur Sentences:` The total number of abstracts with co-occurrences for this term.
4. `Group:` The clustering group that this term was assigned to (only for the "Groups" report format).
5. `List:` The analysis list this term was assigned to (only for the "Lists" report format).
6. `Term:` The query term.
7. `Tags:` The user entered tags for the term.

```
-------------------------------------------------------------------------------
 Total Cooccur Cooccur
  Abs   Terms Sentences Group   List    Term                              Tags
-------------------------------------------------------------------------------
 8942      3    115      1      1  "Acute Myeloid Leukemia" OR "Acute!  [[disease;]]
 1315      1     99      1      1  CD33                                 [[up;]]
  819      1      1      1      1  "leptin receptor"                    [[enr;]]
  162      1      9      2      3  MAD2                                 [[similar;]]
  144      0      0      3      3  Adipsin                              [[down;]]
   97      0      0      5      3  Zyxin                                [[down;]]
   66      1     15      1      1  HoxA9 OR "Hox A9"
   51      1      9      2      3  MAD3
    0                    4      4  Nohitsforthisterm
-------------------------------------------------------------------------------
 11,596 Total abstracts for these terms.
```

QUERY TERMS AS ENTERED

The queries as they were entered on the PDQ_MED input page are shown here. If needed, they can be cut-and-pasted into a new PDQ_MED analysis input page.

```
## sample dataset
[[disease]] "Acute Myeloid Leukemia" "Acute Myelogenous Leukemia"
HoxA9 "Hox A9"
[[up]] CD33
[[enr]]"leptin receptor"
[[down]]Zyxin
MAD3
[[similar]]MAD2
[[down]] Adipsin
Nohitsforthisterm
```

PHARMA TERMS

The pharma terms as they were entered on the PDQ_MED input page are shown here along with the actual pattern match used for the searches. The "Term" column also shows the Perl style regular expression used for pattern matching. The "#" column gives the number of times a particular pharma term occurred with a query term.

As entered:	Term	#
antagonis* agonis* inhibit inhibit* interact* up-regulat* "up regulat*" down-regulat* "down regulat*" regulat* therapy therapies therapeutic drug target target*	inhibit.*?\W	178
	regulat.*?\W	171
	interact.*?\W	156
	therapy	151
	target.*?\W	62
	down[-]*regulat.*?\W	39
	up[-]*regulat.*?\W	33
	therapeutic	33
	drug	33
	therapies	9
	antagonis.*?\W	8
	agonis.*?\W	3
	Total	**1018**

• Hints, Suggestions & FAQ

How Long Does it Take?

Typically, it takes between 0.3 and 1 seconds for every MEDLINE search. The total number of searches, without proximity, is equal to the number of query terms (N). The number of searches with proximity is dependent on the queries. Typical data sets have, on average, two or three proximity neighbors per query. This suggest that, in general, N query terms require ~3N MEDLINE searches.

N	Estimated Time no proximity checking	Estimated Time with proximity checking	Estimated Time with "pharma" checking
10	20 sec.	1 min.	8 min.
100	3 min.	10 min.	1.3 hrs.
1000	30 min.	2 hrs.	13 hrs.

Printing Results

You can use your browser's "Print" command to print your results. Most browsers will not print the graph display correctly. To capture the graph display for printing (or exporting to another program) scroll the **PDQ_MED** output so that the graph display is visible and use a screen capture utility to take a "picture" of the display. For example, on a Windows system use <alt><Print Screen> to take a snapshot of the active window. Open a word processor or image-processing program and paste the captured image into it.

Saving Results

If you would like to save the results of a **PDQ_MED** run use your browsers "Save as ..." command. Make sure to use a file type of .htm or .html. You can now re-open the output file in a web browser, HTML editor or HTML enabled word processing program (such as Microsoft Word). In an HTML editor or Word, you can edit the file. Note that the graph display may not be correctly included in the output file.

Quick Proximity using Title Searching:

For a quick search of MEDLINE, use the MEDLINE Search Field option to restrict the search to "Title Words". This will identify abstracts that contain two (or more) of your query phrases in the *title* of the abstract. Any abstract that contains two of your query terms in the title has a high probability of being of interest. This search has the advantage of being very fast and only retrieving highly relevant abstracts but has the disadvantage of lowered sensitivity.

Known & Potential Problems:

NETWORK ERRORS

If a network error occurs (nothing is returned for the query or the browser times out) then the query is retried for up to 3 hours. If no response is obtained from the web in that time then the whole process aborts.

GRAPH DISPLAY IS TOO COMPLEX

If your query set returns a very large number of co-occurrences in can be difficult to sort out what is significant. A few suggestions to simplify the output;

- Make sure that you have quoted any multiword queries -- "`acute myelogenous leukemia`" is searched as a phrase but `acute myelogenous leukemia` is searched as "`acute`" OR "`myelogenous`" OR "`leukemia`".
- Do a "Title Only" search.
- Set the Grouping Type to RAW and increase the Grouping Cutoff to > 1.
- Set the Grouping Type to WEIGHTED and increase the Grouping Cutoff to > 0.001.
- Use the "Resubmit Graph" button on the output page to change the number of links each node can create.

GRAPH DISPLAY DRIFTS

If the Graph display continues to drift or tumble even after the "energy" has settled down try fixing one or more of the nodes by left-clicking on a node with the mouse, then while still left-clicking also do a right-click. Repeating the mouse click will "un-fix" individual nodes. Note that when any nodes are fixed the auto scaling and centering of the graph is disabled.

4. graphInPharmix Users Guide

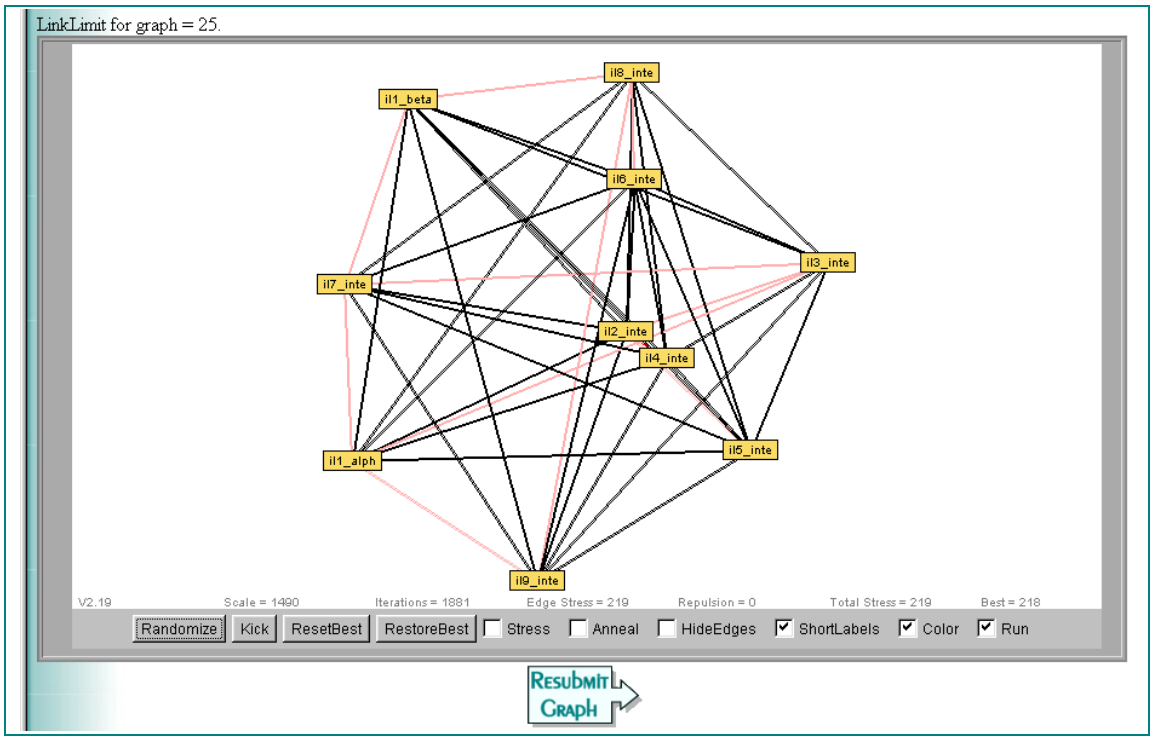
• Description

GraphInPharmix allows the user to graphically re-analyze a data set generated by **PDQ_MED.** Filtering and transform options assist in exploring the interrelationships within the data set. A distance geometry algorithm is used to solve for a two-dimensional representation of linkage data. **GraphInPharmix** provides an interface to the same dynamic Java Applet used by **PDQ_MED**.

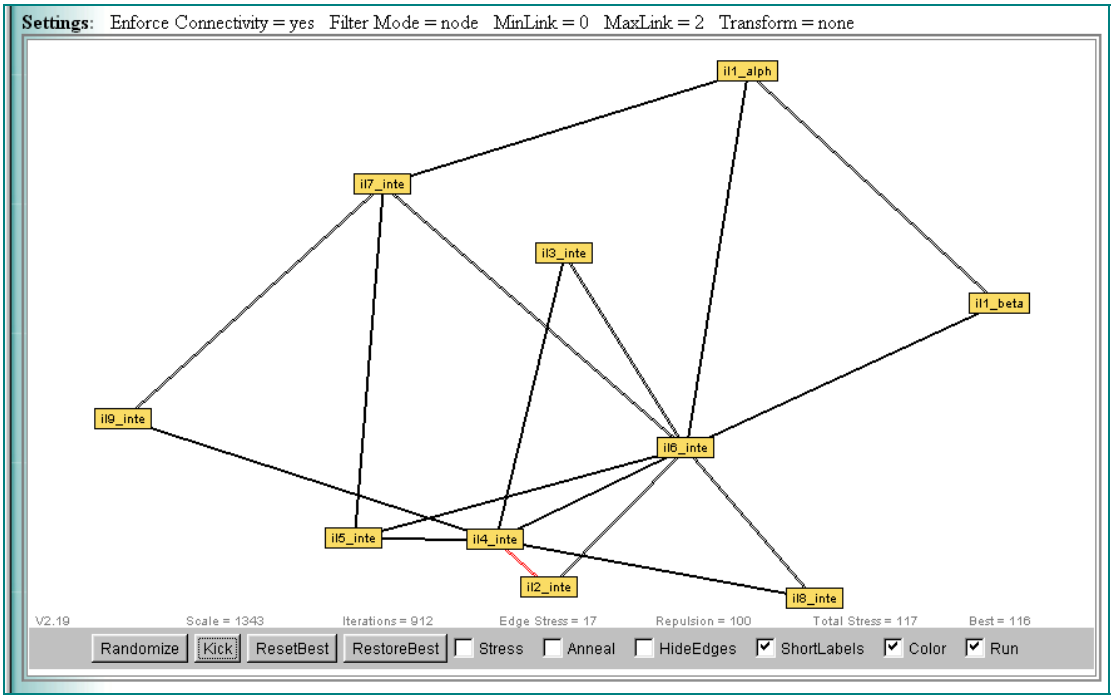
As an example, consider the graph generated by **PDQ_MED** for the query set consisting of the interleukins 1 thru 9;

```
IL1-alpha "interleukin 1alpha" IL1a
IL1-beta "interleukin 1beta" IL1b
IL2 "interleukin 2" IL-2
IL3 "interleukin 3" IL-3
IL4 "interleukin 4" IL-4
IL5 "interleukin 5" IL-5
IL6 "interleukin 6" IL-6
IL7 "interleukin 7" IL-7
IL8 "interleukin 8" IL-8
IL9 "interleukin 9" IL-9
```

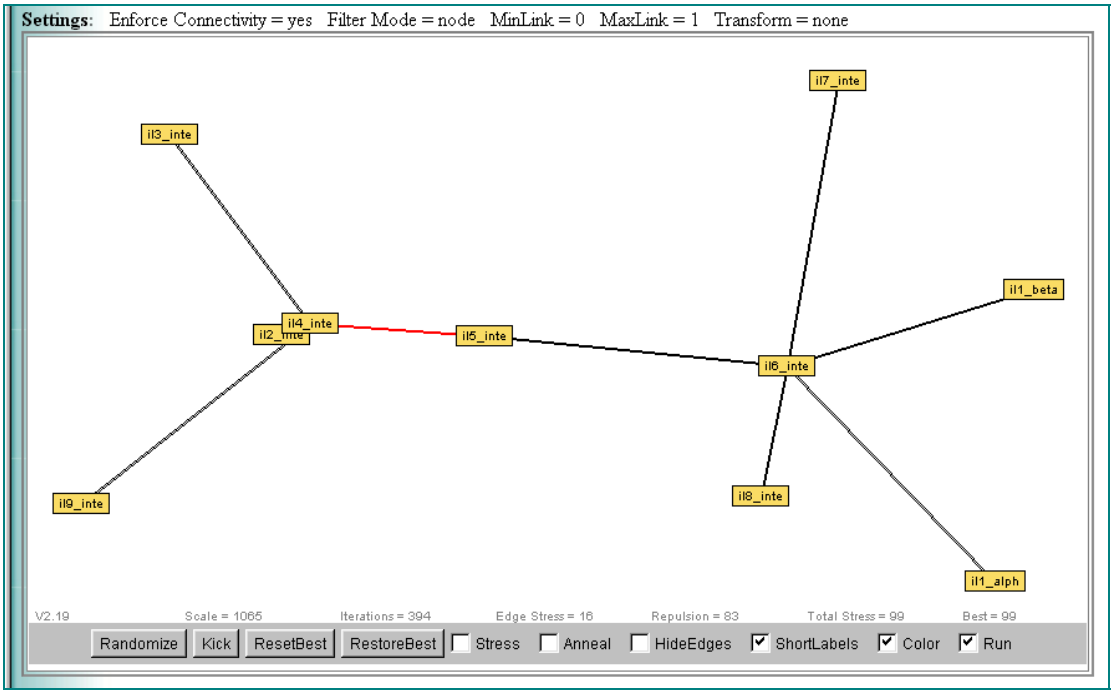
Which gives the highly cross-linked graph;



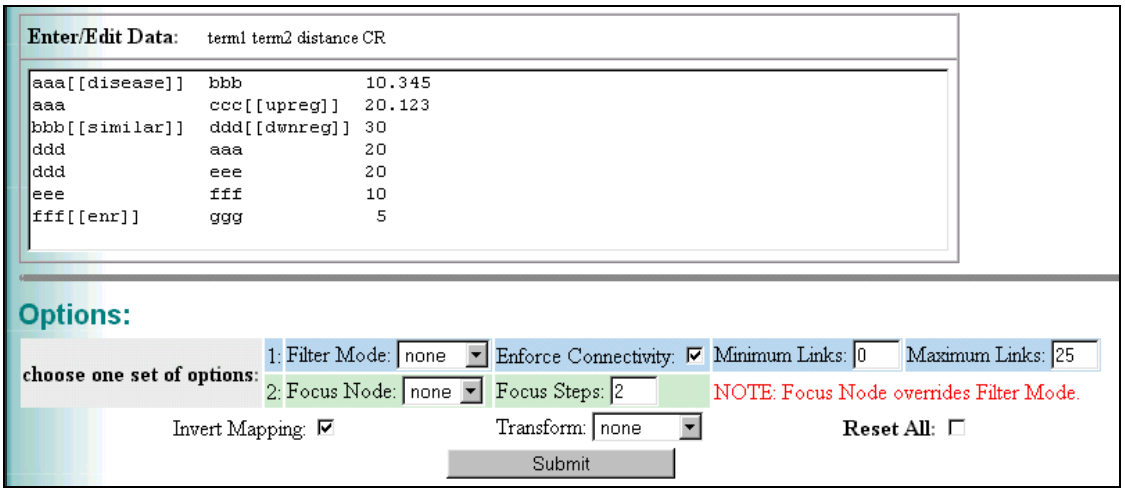
GraphInPharmix allows the user to limit the graph to only the two strongest links from each node. Setting *Minimum Links* to zero and *Maximum Links* to two, directs **GraphInPharmix** to produce the simplified graph;



Limiting the graph to only the strongest link from each node, by setting *Minimum Links* to zero and *Maximum Links* to one, produces the greatly simplified "minimal spanning" graph;



• Input



Data format

The data set is entered in the main text box ("Enter/Edit Data") automatically if **graphInPharmix** is launched from the **PDQ_MED** output page. Alternatively, it can be launched within a browser using the URL;

```
http://localhost/InPharmix/PDQ_MED/graphJPS_interface.pl
```

Where "localhost" is your local path to the InPharmix directory.

The data format is one graph edge per line, white space delimited, formatted as "term1" "term2" and the target length of the connecting edge. Terms containing spaces must be quoted and only letters, numbers and underscores are allowed in the terms.

By default, **graphInPharmix** inverts the target lengths, small numbers lead to long edges and large numbers lead to small edges. This is done since **graphInPharmix** is usually dealing with co-occurrence data and a large number of co-occurrences are displayed with a short edge. This inversion can be turned off with the "Invert Mapping" check box.

• Options

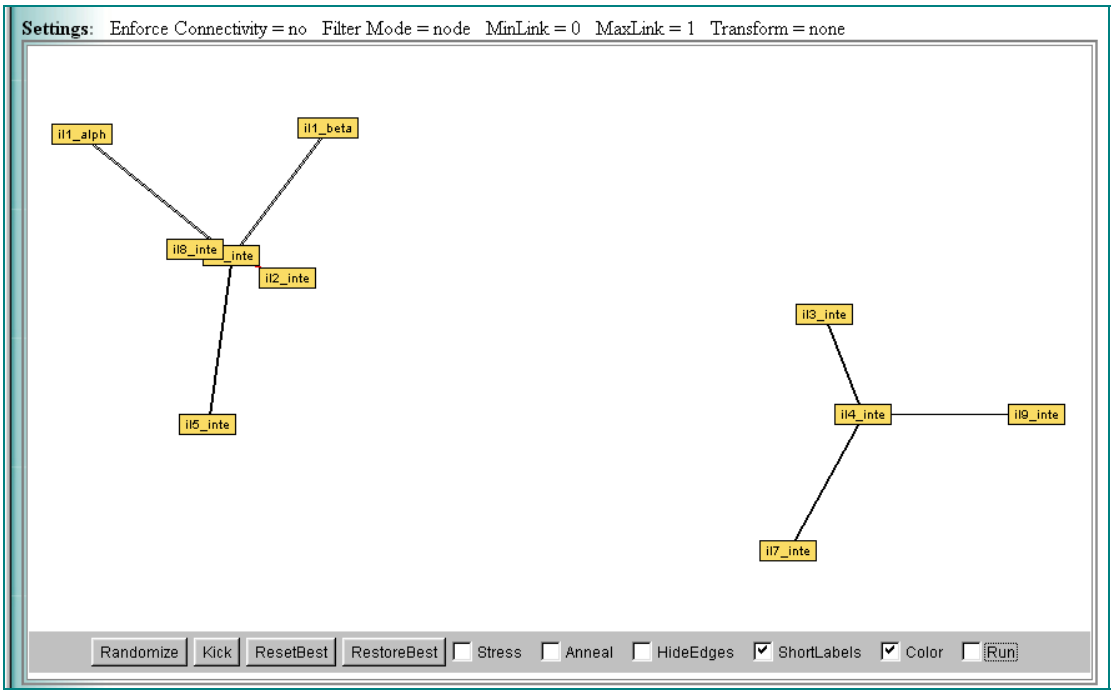
GraphInPharmix provides two methods for simplifying the graph display. Type 1, "Filter Mode," removes edges based on their strength. Type 2, "Focus Node," removes nodes and edges based on the distance from a focus node.

Filter Mode

The Filter Mode can remove the weakest or strongest links from the nodes. If *Filter Mode* is set to "node" then for each node, the minimum and maximum numbers of links are used. For example, if Minimum Links=1 and Maximum Links=2 (see below) then for each node, the two strongest links and the weakest link is used. If *Filter Mode* is "global" then the minimum and maximum numbers apply to the set as a whole. In this case, if minimum=1 and maximum=2, then ONLY the two strongest and one weakest link for the entire set is used.

ENFORCE CONNECTIVITY

The *Filter Mode* option may result in converting what was a single graph into two or more disconnected graphs. For example, using the interleukins sample data with *Maximum Links* set to one, produces the graph;



To force creation of a single graph the *Enforce Connectivity* option may be used. This option, which is on by default, adds back in the minimal set of strongest links required to create a single graph. A useful approach for generating the simplest possible graph is to set *minimum links* and *maximum links* both to zero and use *Enforce Connectivity* to create the graph. This will create the simplest possible graph using the minimal set of strong links.

MINIMUM LINKS

The *number* of links to be used counting from the weakest link. If *Filter Mode*=node, then the minimum links for each node are used. If *Filter Mode*=global, then the minimum links for the entire data set are used.

MAXIMUM LINKS

The *number* of links to be used counting from the strongest link. If *Filter Mode*=node, then the maximum links for each node are used. If *Filter Mode*=global, then the maximum links for the entire data set are used.

Focus Node

The Focus Node option allows the user to focus the graph display on a selected node and suppress nodes distant (defined by counting intervening edges) from it.

NODE SELECTION

The dropdown box contains all of the node names. Select the name of the node that you wish to be the focus.

FOCUS STEP

Focus Steps controls how many edges, and connected nodes, counting away from the Focus Node will be displayed. If step=1 then only the focus node and any nodes directly connected to it are displayed. If steps=2 then the focus node, nodes connected directly to the focus node and nodes connected to nodes connected to the focus node are displayed.

Transform

The user may transform the edge data (distances) prior to graphing. Available transforms are; reciprocal (1/X), Log_{10} (Log10) and squared (X^2). The transforms are not cumulative.

Reset All

The "Reset All" check box resets the options to their default settings. This includes clearing the transform and setting all edges and nodes to be displayed.

Submit

The submit button will reload the graphInPharmix page and display the graph with the selected options.

• Output

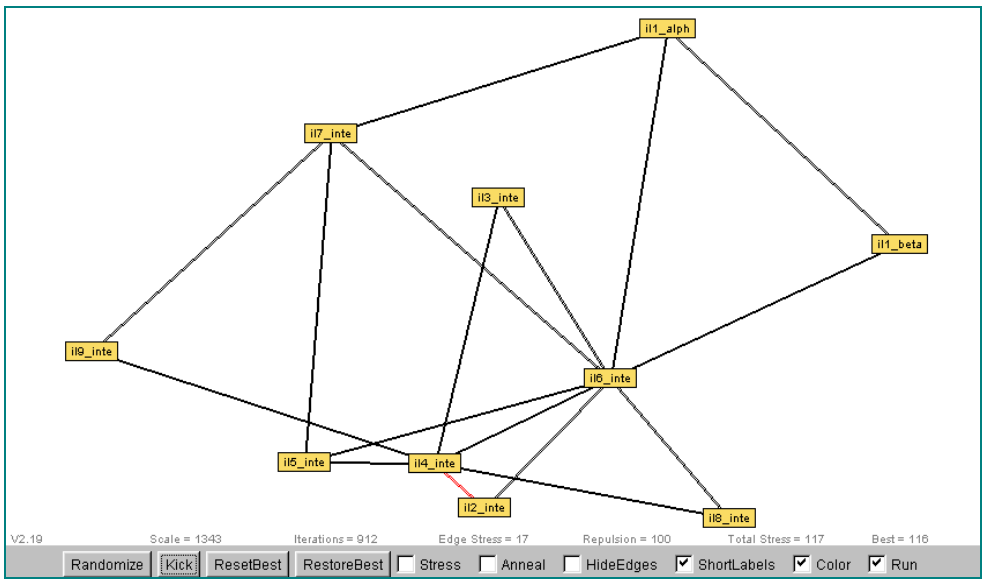
Graph Window

In this graphical analysis of the data, the inverse of the co-occurrence data is used as distance constraints and a solution to this set of constraints in two-dimensional space is searched for. In the graph display, each term (or term phrase) is presented as a box. Terms that co-occur are linked by a line, the target length of which is proportional to the inverse of the co-occurrence frequency (frequent co-occurrences have short lines, infrequent co-occurrences have longer lines). Satisfied edges are shown as black lines, unsatisfied edges as red lines.

There are several controls for this Java Applet to help the user explore their data.

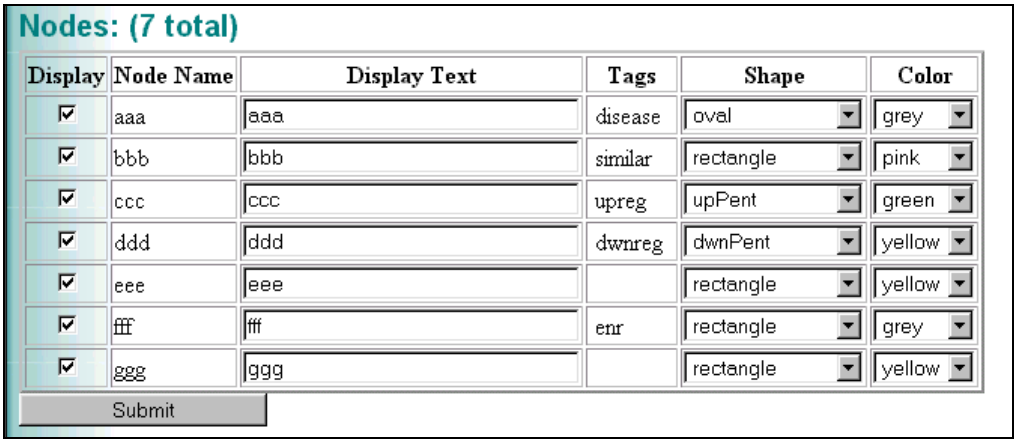
1. Boxes can be repositioned by dragging with the left mouse button.
2. Boxes can be fixed in place, and un-fixed, with the left + right mouse button.
3. *Edge Stress Value:* Displays the total energy contribution from unsatisfied distance (edge) constraints.
4. *Scale:* The current scaling factor for the display with 100 indicating 100%.
5. *Iterations:* The number of iteration (optimization) cycles that have been completed. By default, the java applet stops automatically after 1001 iterations. The optimization can be restarted by clicking twice on the "Run" checkbox.
6. *Edge Stress:* The "energy due" to deviation from the target edge lengths for connected nodes.
7. *Repulsion:* The "energy" contribution from "repulsion", the portion of the energy function which tries to spread the nodes out as much as possible.

8. *Total Stress Value:* Displays the total energy (edge + repulsion) of the current configuration. This is the value the graph algorithm is trying to minimize.
9. *Best:* The best energy value encountered so far. The ResetBest button resets this value.
10. *Randomize Button:* Randomly distributes the nodes in the display window. This is the configuration at the beginning of a minimization run.
11. *Kick Button:* Randomly displaces all nodes by a moderate amount. This button is useful for trying to "untangle" a complex graph.
12. *ResetBest Button:* Resets the Best Energy value.
13. *RestoreBest Button:* Restores the best configuration found since the last Reset.
14. *Stress Checkbox:* Display the stresses associated with each edge. A value of 10 or less is considered satisfied.
15. *Anneal Checkbox:* Initiates a simple "simulated annealing" algorithm to try to locate the optimal configuration for complex graphs. The configuration is periodical "kicked" and then optimized. If a new best configuration is found it is saved. The best configuration encountered can be restored with the *RestoreBest* button.
16. *HideEdges Checkbox:* Hides the display of the lines (edges).
17. *ShortLabels Checkbox:* Toggles between the display of the full term names in the boxes vs. only the first ten characters of each name.
18. *Color Checkbox:* Toggles the display between color and black and white.
19. *Run Checkbox:* Starts and stops the minimizer. When the minimizer is stopped the lines showing the iteration count and energy terms is not shown.



Nodes Table

The "Nodes Table" displays the current settings for each node. The check box in the first column turns the display of the node on and off. The second column gives the name of the node, as it was originally input. The third column allows the name that is actually displayed to be edited. The "Tags" column lists any tags that were in the original data. The "Shape" and "Color" drop down menus change how the node is displayed, overriding the default settings associated with the tags.



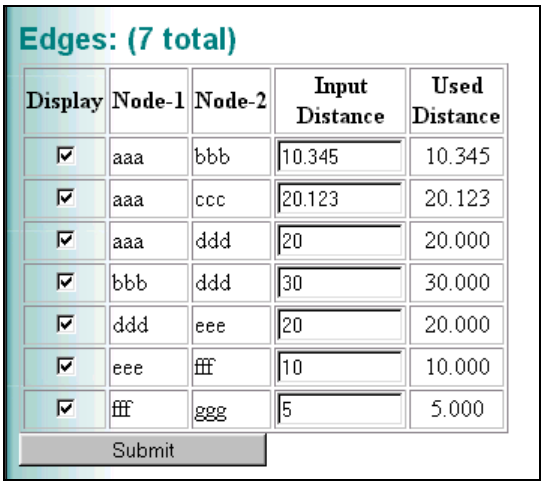
Nodes: (7 total)

Display	Node Name	Display Text	Tags	Shape	Color
☑	aaa	aaa	disease	oval	grey
☑	bbb	bbb	similar	rectangle	pink
☑	ccc	ccc	upreg	upPent	green
☑	ddd	ddd	dwnreg	dwnPent	yellow
☑	eee	eee		rectangle	yellow
☑	fff	fff	enr	rectangle	grey
☑	ggg	ggg		rectangle	yellow

Submit

Data Table (Edges)

The Edges table gives the input term pairs in the second and third columns, the input distance in the fourth column and the value actually used in the graph in the final column. If the edge is not used, then "omitted" is displayed as the used value. If the edge was included to satisfy the Enforce Connectivity option then the value is red. The check box in the first column can be used to turn the display of the edge on or off. The "Input Distance" column is editable.



Edges: (7 total)

Display	Node-1	Node-2	Input Distance	Used Distance
☑	aaa	bbb	10.345	10.345
☑	aaa	ccc	20.123	20.123
☑	aaa	ddd	20	20.000
☑	bbb	ddd	30	30.000
☑	ddd	eee	20	20.000
☑	eee	fff	10	10.000
☑	fff	ggg	5	5.000

Submit

5. Check_Queries Users Guide

• Description:

Check_Queries vets (checks) a set of queries, in **PDQ_MED** format, for proper search parsing by MEDLINE. The queries can include gene and gene product names, diseases, biological processes and other terms. Please note that **Check_Queries** cannot check if the gene name query matches only the gene the user is interested in, instead it just checks that MEDLINE parsed the query correctly.

MEDLINE employs a rather complex parsing algorithm for query terms. For quoted phrases, such as "estrogen receptor", MEDLINE first looks in a phrase index (which includes the MeSH system). If the phrase is not in the phrase index then the individual terms are joined by boolean AND and searched. For example, if the query phrase "estrogen receptor like protein" is not found in the phrase index, then it is searched as;

```
"estrogen[all fields] AND receptor[all fields] AND
    like[all fields] AND protein[all fields]"
```

which does not require that the terms occur consecutively in the document. This re-phrased query will return any document that has these four words, in any order, in any location within the MEDLINE record. In general, this is not what is intended.

A less obvious consequence of this method of handling phrases arises in unexpected places. Unbeknownst to most users, MEDLINE treats hyphens (-) and spaces as equivalent. This is helpful in situations where the common usage of a phrase has both a hyphenated and non-hyphenated version. For example, some authors use "estrogen receptor like protein" whereas others use "estrogen receptor-like protein". In this case, the hyphen is clearly equivalent to a space. In addition, acronyms for proteins are frequently used both with and without a hyphen, for example IL1 and IL-1. However, if a hyphenated version does not occur in MEDLINE then the hyphen-to-space conversion followed by phrase indexing takes place. Consider, for example, the query "CDC20" for which it might be expected, based on the IL1 and IL-1 example, that "CDC-20" would also be an acceptable name. If we query MEDLINE with "CDC20 OR CDC-20" unexpected things happen. In this case, "CDC20" occurs in MEDLINE but "CDC-20" does not. The search engine then converts the hyphen to a space and looks in the phrase index. "CDC 20" does not occur in the phrase index so MEDLINE tries "CDC AND 20" for which MEDLINE now uses the MeSH system, converting the original query to;

```
"CDC20 OR (((CDC[all fields]) OR (Center For Disease
    Control[all fields])) AND 20[all fields])"
```

There are 827 abstracts in MEDLINE that fulfill this search criteria even though CDC20 only occurs in 138 abstracts and CDC-20 does not occur at all. In this case, a reasonable set of acronyms (CDC20 OR CDC-20) gives rise to completely unexpected search results.

For the reasons described above we believe it is important that names be vetted against MEDLINE. By vetting, we mean that each name is searched in MEDLINE and the way in which MEDLINE parses the name is examined to ensure that it is being searched in the desired manner. Any name that is incorrectly searched should be removed from the database.

In addition, for efficiency reasons, any name that is not found in MEDLINE should also be removed from the name database.

• Input

Data format:

The query may be entered in the main text box or uploaded from a local file. In both cases, the format is the same.

Protect spaces in phrases with quote marks, e.g. "interleukin 1", otherwise words and phrases separated by spaces will be searched as individual queries.

Enter the data so that each line contains the name and pseudonyms for a single gene, protein or concept. For example, to vet terms for the estrogen receptor a user might enter the following term list.

```
er1 esra esr1 "estrogen receptor alpha" "estrogen receptor 1"
    "oestrogen receptor"
```

It is important to remember that pseudonyms and acronyms are not necessarily unique to a single gene or protein. Frequently acronyms are common English words that can lead to irrelevant linkages.

MEDLINE Search Field

Check_Queries supports many of the MEDLINE/PUBMED search fields. For a complete description of the PUBMED query system, visit PUBMED. Some of the supported search fields are;

PUBMED Field Name	Description	Notes
ALL	All	default
WORD	Text Word	
TITL	Title Word	
TIAB	Title/Abstract Word	
SUBS	Substance Name	
MAJR	MeSH Major Topic	see Mesh Browser
MESH	MeSH Terms	see Mesh Browser

The Search Field option is global and applies to all of the query terms.

• Example

As an example, consider the following set of queries;

```
"Acute Myeloid Leukemia" "Acute Myelogenous Leukemia" "Acute Myelogenos Leukemia"
cdc20 cdc-20 "cdc 20" "cdc-20"
er1 esra esr1 "estrogen receptor alpha" "estrogen receptor 1" "estrogen receptor" "oestrogen receptor"
nohitsforthisterm
```

Note the misspelling for the "Acute Myelogenos Leukemia" query. For this query set **Check_Queries** responds with;

Note: When viewing the hyperlinks, use your browsers "View Source" option after the page has been loaded to examine the MEDLINE parsing.

```
--------------------------------------------------------------------------------
"Acute Myeloid Leukemia" OR "Acute Myelogenous Leukemia" OR "Acute Myelogenos Leukemia"
   Acute Myeloid Leukemia          (Search)count=   5138   OK
   Acute Myelogenous Leukemia      (Search)count=   3026   OK
   Acute Myelogenos Leukemia       (Search)count=  62203   QUERY NOT FOUND (Phrase not found)(Quoted Phrase
not found)
      original query  = 62203 hits
      corrected query = 8032 hits
--------------------------------------------------------------------------------
cdc20 OR cdc-20 OR "cdc 20" OR "cdc-20"
   cdc20                           (Search)count=    144   MAPPED TO: ("p55CDC protein"[Substance Name] OR
cdc20[Text Word])
   cdc-20                          (Search)count=      0   BAD MAPPING TO: ("centers for disease control
and prevention (u.s.)"[MeSH Terms] OR cdc[Text Word])
   cdc 20                          (Search)count=      0   QUERY NOT FOUND (Quoted Phrase not found)
   cdc-20                          (Search)count=      0   BAD MAPPING TO: ("centers for disease control
and prevention (u.s.)"[MeSH Terms] OR cdc[Text Word])
      original query  = 851 hits
      corrected query = 144 hits
--------------------------------------------------------------------------------
er1 OR esra OR esr1 OR "estrogen receptor alpha" OR "estrogen receptor 1" OR "estrogen receptor" OR
"oestrogen receptor"
   er1                             (Search)count=     34   OK
   esra                            (Search)count=     13   OK
   esr1                            (Search)count=    110   OK
   estrogen receptor alpha         (Search)count=   1385   OK
   estrogen receptor 1             (Search)count=      0   QUERY NOT FOUND (Quoted Phrase not found)
   estrogen receptor               (Search)count=  12408   OK
   oestrogen receptor              (Search)count=   1900   OK
      original query  = 17021 hits
      corrected query = 14276 hits
--------------------------------------------------------------------------------
```

For the "Acute Myelogenous Leukemia" queries, the first two versions occur in MEDLINE's phrase index and are searched properly. The third sub-query, "Acute Myelogenos Leukemia", is misspelled giving rise to a very large number of matches in MEDLINE (62,203) since it was not searched by MEDLINE as a phrase but instead as "acute AND leukemia".

In the second set of queries, CDC20 is mapped by the MeSH system to what appears to be a correct name (p55CDC protein). All three of the remaining queries do not occur in the phrase index and so MEDLINE searched them as ANDed terms. Check_Queries has removed these terms. Note that the hit count for the original query (851) does not match the sum of the individual queries. This appears to be a bug in the MEDLINE search engine.

For the final set of queries, concerning the estrogen receptor, all of the terms appear to parse correctly with the exception of "estrogen receptor 1" which happens to be the current accepted name as defined by HUGO.

The next portion of the **Check_Queries** output contains a list of corrected terms suitable for us with **PDQ_MED** as well as listings of complete queries (entire input lines) that were removed and sub-queries that were removed.

Corrected Queries (vetted for MEDLINE field: "ALL")

```
"Acute Myeloid Leukemia" "Acute Myelogenous Leukemia"
cdc20
er1 esra esr1 "estrogen receptor alpha" "estrogen receptor"
"oestrogen receptor"
```

NOTE: These query LINES were completely removed!

```
nohitsforthisterm
```

Bad / Removed Sub-Queries

```
"Acute Myelogenos Leukemia"
cdc-20
"cdc 20"
"cdc-20"
"estrogen receptor 1"
nohitsforthisterm
```

6. Namer Users Guide

• Introduction:

InPharmix has generated databases of gene and gene product names suitable for searching in MEDLINE. **Namer** provides a method of retrieving names from these databases for a set of microarray gene identifiers. Note that **Namer** does not actually generate the names. Instead, it provides an interface to name lists that InPharmix has already generated.

The output from **Namer** can be used as the input for **PDQ_MED**.

Background

New techniques in genomics research allow scientists to query the expression levels of thousands of genes in a single experiment. The massive amount of data that results from these types of experiments presents a new set of challenges for knowledge extraction that did not exist a few years ago.

One critically important resource for identifying actionable information in gene lists is the scientific literature. The scientific literature is the central repository of biological knowledge. Biologists rely on access to the literature to identify what is already known, and build on this existing framework of knowledge with their own experiments. Currently, the most widely electronically accessible repository of the biomedical literature is MEDLINE, which contains over 11 million abstracts (more than two billion words) and is growing rapidly.

Unfortunately, the interoperability between gene lists, sequence databases and the scientific literature is poor. It is surprisingly difficult to take a gene name found in a nucleotide sequence database record and effectively search the biomedical literature. This is a different problem than the one being addressed by HUGO and other gene ontology efforts. Here we must deal not only with the accepted names but also with the legacy data of the scientific literature.

Naming Difficulties Examples

Consider, for example, the nucleotide sequence databases. GENBANK, the most commonly used, does not contain a "gene product" name field. Instead, the name is imbedded in other information. For example, the GENBANK nucleotide definition for "Estrogen Receptor 1" (the HUGO accepted name for this gene) is;

```
DEFINITION  Homo sapiens estrogen receptor 1 (ESR1), mRNA.
```

This phrase is unsuitable for searching in MEDLINE. Indeed, if the entire phrase is submitted as a query to MEDLINE, only two abstracts are found. The user must recognize that the name of the gene product is "Estrogen Receptor 1", that ESR1 is an acronym and the reset of the information is extraneous. If the query is submitted to MEDLINE as;

```
"estrogen receptor 1" OR ESR1
```

then only 12 abstracts are found and MEDLINE complains "Quoted phrase not found". Clearly, there are more than 12 abstracts concerning this receptor in MEDLINE. The GENBANK record gives no indication that ESR1 was previously know as "Estrogen Receptor" (prior to 1998) and as "Estrogen receptor alpha" (1998-2001). If these older names are recognized and included in the query, then more than 20,000 abstracts are found in MEDLINE.

The next issues that need to be addressed are the difference between the name of a gene product and the name of the biologically active entity. We can identify two distinct types of naming: "Included names" and "Holo names".

Included Names

Consider a protein sequence that is cleaved to shorter, active peptides. For example, complement C3 precursor is processed into anaphylatoxin (c3a) and c3b. When trying to associate the affects of changes in expression (e.g., from a microarray experiment) with biological response the researcher should examine the literature for the active forms of the gene product. In this case, the literature concerning complement C3, anaphylatoxin, c3a and c3b should be explored.

HOLO Names

Another variation of the naming problem is "Holo names", the names of the active biological species for heterogeneous protein complexes. Examples would include FOS + JUN forming AP-1, HLA + beta-2 microglobulin forming the class I MHC antigen, and IgG heavy chain + IgG light chain forming an antibody. As in the case of included names, when examining the possible ramifications of changes in the expression of a gene it is necessary to also explore the "holo"-proteins of which they are a part of. For example, if FOS is found to be regulated, then the literature concerning both FOS and AP-1 may contain relevant information about the biological ramifications of the regulation of FOS. In these cases, it is clear that the "name" used to search the literature must take into account more than just the simple gene product name.

Name Vetting

A final issue that needs to be addressed is how MEDLINE actually parses and responds to queries. MEDLINE employs a rather complex parsing algorithm for query terms. For quoted phrases, such as "estrogen receptor", MEDLINE first looks in a phrase index (which includes the MeSH system). If the phrase is not in the phrase index then the individual terms are joined by boolean AND and searched. For example, if the query phrase "estrogen receptor like protein" is not found in the phrase index, then it is searched as;

```
"estrogen[all fields] AND receptor[all fields] AND
    like[all fields] AND protein[all fields]"
```

which does not require that the terms occur consecutively in the document. This re-phrased query will return any document that has these four words, in any order, in any location within the MEDLINE record. This is clearly not what was intended.

A less obvious consequence of this method of handling phrases arises in unexpected places. Unbeknownst to most users, MEDLINE treats hyphens (-) and spaces as equivalent. This is helpful in situations where the common usage of a phrase has both a hyphenated and non-hyphenated version. For example, some authors use "estrogen receptor like protein" whereas others use "estrogen receptor-like protein". In this case, the hyphen is clearly equivalent to a space. In addition, acronyms for proteins are frequently used both with and without a hyphen, for example IL1 and IL-1. However, if a hyphenated version does not occur in MEDLINE then the hyphen-to-space conversion followed by phrase indexing takes place. Consider, for example, the query "CDC20" for which it might be expected, based on the IL1 and IL-1 example, that "CDC-20" would also be an acceptable name. If we query MEDLINE with

"CDC20 OR CDC-20" unexpected things happen. In this case, "CDC20" occurs in MEDLINE but "CDC-20" does not. The search engine then converts the hyphen to a space and looks in the phrase index. "CDC 20" does not occur in the phrase index so MEDLINE tries "CDC AND 20" for which MEDLINE now uses the MeSH system, converting the original query to;

```
"CDC20 OR (((CDC[all fields]) OR (Center For Disease Control
   [all fields])) AND 20[all fields])"
```

There are 827 abstracts in MEDLINE that fulfill this search criteria even though CDC20 only occurs in 138 abstracts and CDC-20 does not occur at all. In this case then, a reasonable set of acronyms (CDC20 OR CDC-20) gives rise to completely unexpected search results.

For the reasons described above we believe it is important that names be vetted against MEDLINE. By vetting, we mean that each name is searched in MEDLINE and the way in which MEDLINE parses the name is examined to ensure that it is being searched in the desired manner. Any name that is incorrectly searched should be removed from the database. In addition, for efficiency reasons, any name that is not found in MEDLINE should also be removed from the name database.

Clearly then, there are significant "interoperability" issues between GENBANK and MEDLINE. Some of the issues outlined above can be addressed by using information available in other database. For human genes, GeneCards, Online Mendelian Inheritance in Man (OMIM), the Genome Database (GDB) and the Human Genome Organization (HUGO) databases can provide additional information. However, manually extracting information from these multiple data sources for large sets of genes is impractical.

- # Input:

Query Format

The input to **Namer** can be via a file or by pasting or typing entries in the query box. The query format is simply a list of sequence IDs or GENBANK GI numbers separated by spaces, tabs or line breaks For example, for an Affymetrix chip;

```
L47345_at
X99268_at   X86012_at
AB000410_s_at
```

For an array that uses GENBANK IDs , the format is;

```
AI044326
AI044423
AI044424
AI044452
```

Database Format

The database flat file that **Namer** uses is tab delimited text file. The four tab delimited columns are;

1. Sequence or spot ID number from the microarray.
2. GENBANK GI number.
3. A semicolon delimited list of standard names, pseudonyms, acronyms and aliases for the gene.
4. A semicolon delimited list of holo and included names, if any, for the gene.

For some databases, the sequence ID and GENBANK GI number may be the same. The users queries can be expressed as sequence IDs or GENBANK IDs.

Microarray Name:

Use the drop down menu to select the DNA microarray that is the source of your genes. InPharmix is in the process of naming other microarrays from a variety of vendors. If there is a particular array that you would like names for please let us know.

• Options

Include holo/included Names:

If this option is selected, **Namer** provides both the standard names and acronyms as well as names for any included or holo products. The default is to include holo/included names.

• Description of the Output:

The first segment of the output gives the **Namer** version number, run title, date and time.

```
Project type: Namer
Version: 0.20, 15 August 2002
Host Site: InPharmix Inc. This is a full license for
InPharmix Inc. software which expires 31 Dec 2002.
```

The next segment of the output provides information about the number of IDs the user has entered as well as information about the names database file that is being used.

```
Total IDs to process = 12
7070 entries found in the database flat file
(Namer_HU6800.Namer_database)
Database Dated: Aug 16, 2002 at 23:04.29 .
Database Comments: This is a preliminary release of this
dataset.
```

Namer now lists the names found for each of the input IDs. The IDs themselves are not shown. Blank lines indicate where in the input list sequences occurred for which there are no useable names in the database. These names are correctly formatted for use as input for **PDQ_MED**.

The names for your ID's --
See also the output files listed below.

```
MLR NR3C2 NR3C-2 "aldosterone receptor" "mineralocorticoid receptor"
CBX5 CBX-5 HP1HS-ALPHA

SIII TCEB3 TCEB-3 "transcription elongation factor B"
OAS2 OAS-2 "P69 2-5A synthetase" "2'-5'oligoadenylate synthetase 2"
GC-F GUCY2F GUC2DL RetGC-2 retina-specific "retinal guanylate cyclase 2F"

MIN3 CD59 MIN2 MIN1 MIN-1 MIN-2 MSK21 CD-59 MSK-21 MIC-11 "CD59 antigen p18-20"
F8C FVIII "Factor VIII" "procoagulant component coagulation factor VIIIc"
```

The final segment of the output provides information on the number of IDs that were named, the number that were found but have no useable names, and the number that were not found in the database file. In addition, links are provided to two output files. The first file contains both the sequence IDs and the names with the sequence ID separated from the names by a tab character. The second file contains just the names for each sequence and is suitable for use as the input to **PDQ_MED**.

OK All 12 ID's were found in the database file.
But, 2 ID's do not have any useable names. These ID's were;

```
AB000466_at
U90911_at
```

Writing to output files ... Done.
The output files are available in two formats,
sequence IDs + names and names only.

7. StAT Users Guide

• Description

StAT (**St**atistical **A**nalysis of **T**ext) is a statistical and heuristic based tool that allows the researcher to quickly explore the literature concerning a single topic or set of abstracts. **StAT** can pass a user query to MEDLINE or accept pasted or uploaded abstract files directly. **StAT** then processes the set of returned/entered abstracts, highlighting important concepts in the literature set. **StAT** has an integrated MEDLINE interface, a pharma-specific knowledge domain and a user-friendly interface.

Method

The **StAT** Processor begins with a set of abstracts (returned from MEDLINE or input by the user), strips out *common words*, reduces plurals, marks *"Pharma" terms* (optional), and compares the frequency of the remaining terms of the test set to the frequency of those terms in the *background set*. Terms which appear with significantly greater frequency in the test set relative to the background set are marked as "key" terms. Terms that do not occur in the background set and terms in the "Pharma" list are also marked as "key" terms. **StAT** then ranks the key terms, sentences and abstracts, and presents these ordered lists to the user.

Background Sets

The **StAT** Processor compares the word frequencies in the input abstracts to a large set of background abstracts. The size of the background set ensures statistical significance. The domain of the background set controls relevance ranking.

Results

The results of the **StAT** Processor are returned to the user in a format designed to help the user quickly identify the significant details pertaining to the query set, and to focus on the abstracts most likely to include interesting or relevant information.

• Quick Start

Once installation is completed, you can initiate a **StAT** run by simply entering a query term and selecting "Submit". For example, entering;

```
"Zinc alpha-2 glycoprotein" OR zag
```

searches MEDLINE for abstracts that contain the phrase "zinc alpha-2 glycoprotein" OR abstracts that contain "zag".

Throughout the input and output forms the  icon links to the relevant section in the online help manual.

• Input

The **StAT** Processor requires as input a set of abstracts formatted to include an abstract ID and the abstract text. Abstract titles are not needed. The input for the **StAT** Processor can be generated in one of three ways. You can enter a query to MEDLINE, or paste or upload the abstracts directly to **StAT**. If you enter a query in the query box, on initiating the run, **StAT** will pass the query to MEDLINE, accept the response from MEDLINE, and convert the

response into the format required by the **StAT** Processor. If you choose to upload or paste in the abstracts, they must conform to one of the three formats supported by the **StAT** Processor.

MEDLINE Query Formats

For a complete discussion of MEDLINE query formats, visit the *NCBI/PubMed* (http://www.ncbi.nlm.nih.gov/entrez/) site. Some guidelines for query formats are outlined below.

Entered As:	**Searched As:**
"single cell"	Search a phrase
dna crick	Implied Boolean AND
dna AND crick	Explicit Boolean AND
estrogen OR 17b-estradiol	Boolean OR
infecti*	Wild cards (includes "infection", "infective" ...)

Note: AND, OR, NOT and other boolean functions are case specific. Parenthesis may be used to control the grouping of terms as in;

```
"zinc alpha-2 glycoprotein" OR ( zag NOT zig )
```

DATES & DATE RANGING

Dates or date ranges must be entered using the format YYYY/MM/DD [date field], e.g., 1997/10/06 [edat] or 1998/03/15 [dp]. The month and day are optional, e.g., 1997 [edat] or 1997/03 [dp]. To enter a date range, insert a colon (:) between each date, e.g., 1993:1995 [edat] or 1997/01:1997/06 [edat].

PUBMED'S DATE FIELDS

- Date of Publication [DP]
- Entrez Date [EDAT] -- The date the citation first entered PubMed.
- MeSH Date [MHDA] -- The date the citation was indexed with MeSH terms.

A **StAT** query for MMP-9 restricted to the month of March, 2001 would be;

```
mmp-9 AND 2001/03/01:2001/03/31[dp]
```

For additional field descriptors, see the section below titled "MEDLINE Search Fields".

Manual Formats

If you have elected to paste or upload your own abstract data, you must use one of the three formats described below.

MEDLINE FORMAT

The only required fields for the MEDLINE format are UI, AB and PMID (shown in bold below), all other fields are ignored.

```
UI  - 20247761  <-- Required!
AU  - Sadaka HA
AU  - Allam SR
AU  - Rezk HA
AU  - Abo-El-Nazar SY
AU  - Shola AY
TI  - Isolation of dust mites from houses of Egyptian allergic patients and
      induction of experimental sensitivity by Dermatophagoides pteronyssinus
      [In Process Citation]
LA  - Eng
DA  - 20000428
DP  - 2000 Apr
IS  - 0253-5890
TA  - J Egypt Soc Parasitol
PG  - 263-76
CY  - EGYPT
AB  - Six house dust mite (HDM) species were isolated from dust <-- Required!
      of floors and mattresses of allergic patients houses in Alexandria.
      D. pteronyssinus (D.p.) was the dominant species in dust of floors and
      mattresses with average percentages of 68.9% and 78.3% respectively. It
      was used to induce experimental sensitivity in Swiss albino mice by
      repeated weekly intranasal instillation of D.p. mites in phosphate buffer
      saline (PBS). Cytological examination of bronchoalveolar lavage (BAL)
      fluid of mice revealed prolonged eosinophilia, that peaked on day 28 of
      the experiment and persisted till the end of the study. Blood eosinophilic
      counts were progressively increased during the course of the experiment.
      Histopathological findings showed evident eosinophilic infiltration in
      nasal and lung tissues of the sensitized mice.
AD  - Department of Parasitology, Faculty of Medicine, University Student
      Hospital, Alexandria University, Egypt.
RO  - O:099
PMID- 0010786037  <-- Required!
MHDA- 2000/04/29 09:00
EDAT- 2000/04/29 09:00
SO  - J Egypt Soc Parasitol 2000 Apr;30(1):263-76
```

"FASTA" FORMAT WITH PMIDS

The "Fasta" format with PMIDs (PubMed Identification number) consists of a line beginning with a ">", a valid PMID followed by a newline and the abstract. Subsequent abstracts begin with a line starting with ">" followed by another PMID.

```
>0010786037 <-- Required!
Six house dust mite (HDM) species were isolated from dust of floors and mattresses of
allergic patients houses in Alexandria. Dermatophagoides pteronyssinus (D.p.) was the
dominant species in dust of floors and mattresses with average percentages of 68.9%
and 78.3% respectively. It was used to induce experimental sensitivity in Swiss albino
mice by repeated weekly intranasal instillation of D.p. mites in phosphate buffer
saline (PBS). Cytological examination of bronchoalveolar lavage (BAL) fluid of mice
revealed prolonged eosinophilia, that peaked on day 28 of the experiment and persisted
till the end of the study. Blood eosinophilic counts were progressively increased
during the course of the experiment. Histopathological findings showed evident
eosinophilic infiltration in nasal and lung tissues of the sensitized mice.
```

"FASTA" FORMAT

The "Fasta" format consists of a line beginning with an ">", a unique abstract identifier, followed by a newline and the abstract. Subsequent abstracts begin with a line starting with ">" followed by another identifier.

```
>Abstract1  <-- any unique string (no spaces)
Six house dust mite (HDM) species were isolated from dust of floors and mattresses of
allergic patients houses in Alexandria. Dermatophagoides pteronyssinus (D.p.) was the
dominant species in dust of floors and mattresses with average percentages of 68.9%
and 78.3% respectively. It was used to induce experimental sensitivity in Swiss albino
mice by repeated weekly intranasal instillation of D.p. mites in phosphate buffer
saline (PBS). Cytological examination of bronchoalveolar lavage (BAL) fluid of mice
revealed prolonged eosinophilia, that peaked on day 28 of the experiment and persisted
till the end of the study. Blood eosinophilic counts were progressively increased
during the course of the experiment. Histopathological findings showed evident
eosinophilic infiltration in nasal and lung tissues of the sensitized mice.
```

• Parameters

StAT has several parameters that may be modified by the user. If you do not specify a choice, **StAT** will use default settings.

• MEDLINE Query Options

MEDLINE Search Field

StAT supports many of the MEDLINE/PUBMED search fields. For a complete description of the PUBMED query system, visit PUBMED (`http://www.ncbi.nlm.nih.gov-/entrez/query/static/help/pmhelp.html`). The supported search fields are;

PUBMED Field Name	Description	Notes
ALL	All MEDLINE Fields	default
WORD	Text Word	
TITL	Title Word	
TIAB	Title/Abstract Word	
SUBS	Substance Name	
MAJR	MeSH Major Topic	see Mesh Browser
MESH	MeSH Terms	see Mesh Browser
AUTH	Author Name	
ECNO	EC/RN Number	
EDAT	Entrez Date	
MHDA	MeSH Date	
DP	Publication Date	
JOUR	Journal Name	
PT	Publication Type	
VOL	Volume	

The Search Field option applies to all of the terms in the MEDLINE query box. If you would like to specify a particular search field for just one of your query terms in the MEDLINE query box follow the term with the field name in square brackets. For example;

`"zinc alpha-2 glycoprotein"` **`Brysk[AUTH]`**

Note the implied AND between the quoted phrase and the author name.

LANGUAGE

There are two options for Language; **All** and **English**. This restriction is on the language of the article which may or may not be the same as the language of the abstract. Most non-English entries in MEDLINE have English abstracts.

MAXIMUM NUMBER OF ABSTRACTS

When entering a query to MEDLINE, you may select the maximum number of abstract to be returned to **StAT**. The default value is currently set to 25 abstracts. Note that the greater the number of abstracts returned, the longer the processing time. Also, note that MEDLINE may return fewer abstracts than requested, and that some of the records may not have abstracts. Records without abstracts are ignored by **StAT**.

Upload Local File Format

The only option for an uploaded local file is the file format. Acceptable formats are;

MEDLINE	Standard MEDLINE Format
"Fasta" with PMIDs	"Fasta" format with MEDLINE IDs (PMIDs)
"Fasta"	"Fasta" format

PASTE TEXT FORMAT

NOTE: Most browsers limit the number of characters that can be pasted. Typically, this limit is less than 30K characters (20~30 abstracts).

The only option for pasted text is the data format. Acceptable formats are;

MEDLINE	Standard MEDLINE Format
"Fasta" with PMIDs	"Fasta" format with MEDLINE IDs (PMIDs)
"Fasta"	"Fasta" format

• Background Data Sets

You may select the background data against which your set of abstracts is compared. At this time, **StAT** has three data sets, "Tissues and Diseases", "Genes and Proteins" (GenPro) and "All". The default set is GenPro.

GenPro

The background set "GenPro" contains statistics on all terms that appear in a set of 1300 abstracts pertaining to genes and proteins (13 groups). The set has been stripped of common English words, and plurals have been reduced to singular. Currently, this set contains 11,500 terms.

Tissues and Diseases

The background set "Tissues and Diseases" contains statistics on all terms that appear in a set of 3000 abstracts pertaining to tissue types and diseases, e.g. liver, cancer (30 groups). The set has been stripped of common English words, and plurals have been reduced to singular. Currently, this set contains 19,900 terms.

All

The background set labeled ALL contains statistics on all terms that appear in either of the other two sets. The ALL background set is not merely a concatenation of data for the two sets. Rather, the average frequency of each term was calculated as a member of the entire set. As in the GenPro and Tissue and Disease sets, common English words and plurals have been stripped. Currently, this set contains 22,400 terms.

• Z-Score Cutoff

In the process of evaluating the significance of individual terms, the **StAT** processor generates a Z-score for each term. The Z-score is proportional to the deviation of the frequency of the term in the test set as compared to the frequency of the term in the background set. In other words, the closer the frequency of the term in the test set to the "expected" frequency (the frequency in the background set), the lower the Z-score. The Z-score is used by the **StAT** processor in ranking the significance of terms, sentences and abstracts. Using this Z-score, the **StAT** Processor considers terms that are rare in the background set, but common in the test set, as significant.

There are two types of terms for which **StAT** calculates the Z-score differently. For "new words", words that do not occur in the background set, **StAT** assigns a Z-score of 1/10th the maximum Z-score for the set of words. For "Pharma Terms", **StAT** overrides the calculated Z-score and uses a value of 1/2 the maximum Z-score for the set of words.

You can change the cut-off for significance for your data set by entering your own value for the Z-score option. Although a Z-score is calculated for each term, you can decide which Z-scores indicate a significant term. At this time, the default setting for Z-score is 15. The range for the Z-score in a given set of abstracts will vary depending on the similarity between your set and the background set. The more the two sets have in common, the lower the Z-scores of the test set terms are likely to be. Since the Z-score cutoff determines which terms will be marked, changing the cutoff can change the amount and quality of output, especially in the listing of key-words and ranked sentences.

• Use Pharma Terms

This option allows the user to turn off the use of the "Pharma" terms. The default is to use the Pharma terms.

• Include Titles

If your data is in MEDLINE format, (from a MEDLINE query or pasted or uploaded in MEDLINE format) this option allows the user to include the article titles in the abstract. The default is to use the article titles.

• Output

StAT generates five blocks of output; *Run Statistics*, *Keywords*, *Ranked Sentences* and *Ranked Abstracts*. Links between the summary, words, sentences and abstracts facilitate movement around the results listing. Details of the five blocks of output are given below.

Run Statistics

The first portion of the **StAT** output shows the parameters and various statistics for the run. The first table echoes the user input parameters including the MEDLINE query (if used). The second portion of the run statistics gives the URL of the query sent to MEDLINE.

Options and Parameters

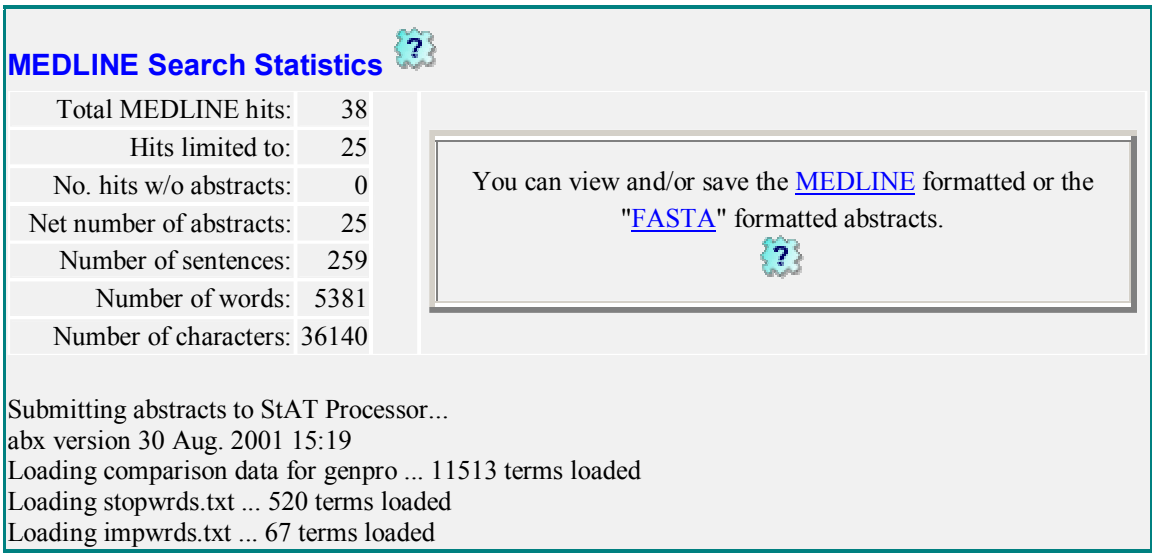
MEDLINE Query:	"zinc alpha-2 glycoprotein" OR (zag NOT zig)
MEDLINE Language:	ALL
MEDLINE Field:	TIAB
Maximum Abs:	25
Z-cutoff:	10
Background Set:	genpro
Use Pharma:	Yes
Use Titles:	Yes

MEDLINE Query URL =
http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=SEARCH&db=PUBMED&doptcmdl=MEDLINE&term="zinc+alpha-2+glycoprotein"[TIAB]+OR+(zag[TIAB]+NOT+zig[TIAB])&dispmax=25

Contacting MEDLINE... Response time: 3 seconds.
Extracting MEDLINE Data...
Reformatting Data...
Abs:1297243 Truncated at 250 words by MEDLINE
Collecting Statistics...

The second segment of the run log lists;

1. The number of matching records found in MEDLINE.
2. The number of hits limited by the *Maximum number of abstracts* parameter.
3. The number of records returned by MEDLINE that do not have abstracts.
4. The net number of abstracts (limited hits minus number missing abstracts).
5. The total number of sentences in the abstracts.
6. The total number of words in the abstracts.
7. The total number of characters in the abstracts.



MEDLINE Search Statistics

Total MEDLINE hits:	38
Hits limited to:	25
No. hits w/o abstracts:	0
Net number of abstracts:	25
Number of sentences:	259
Number of words:	5381
Number of characters:	36140

You can view and/or save the MEDLINE formatted or the "FASTA" formatted abstracts.

Submitting abstracts to StAT Processor...
abx version 30 Aug. 2001 15:19
Loading comparison data for genpro ... 11513 terms loaded
Loading stopwrds.txt ... 520 terms loaded
Loading impwrds.txt ... 67 terms loaded

You may view and/or export the abstracts resulting from a MEDLINE query by clicking on the appropriate link.

Ranked Sentences

The **StAT** Processor uses the calculated Z-score and the user entered cutoff to determine the significance of individual sentences in the abstracts. A Z-score is calculated for each sentence that is the average of the scores of scored terms in the sentence. Not all words in the sentence will contribute to this score, only words with Z-scores will contribute. Therefore, a sentence that has many common English terms (e.g. "the", "a") may have a relatively low Z-score. As with keywords, you can display all sentences, in ranked order, by setting the cutoff to zero.

The sentences are displayed to the user ranked by average Z-score, in order from highest (most significant) to lowest (least significant). Within each sentence, words with a Z-score above the user defined cutoff are highlighted in blue, new words are in green and Pharma terms are in red. Clicking on a highlighted term will jump to that term in the keywords list. In addition, the abstract in which the sentence appears is identified, with a link that allows you to jump to the full abstract to examine the sentence in context. Please note, sentence fragments may occasionally appear due to individual quirks of the abstract in MEDLINE.

StAT Ranked Sentences

Color Key: new term, pharma term, keyword. (Keywords, Abstracts)

Z Score	Abstract Link	Sentence
25.12	11455581	Zinc-alpha(2)-glycoprotein (Znalpha(2)gp) is widely distributed in body fluids and epithelia.
19.79	10462714	Zn-alpha(2)-glycoprotein (Znalpha(2)gp) is a soluble protein widely distributed in body fluids and glandular epithelia.
18.88	9413177	We have also detected it in human stratified epithelia (epidermis and buccal mucosa).
18.75	11205870	TITLE: Dipeptidyl peptidase IV is a target for covalent adduct formation with the acyl glucuronide metabolite of the anti-inflammatory drug zomepirac.
18.73	9413177	Zinc-alpha 2-glycoprotein has been detected in most body fluids, and its antibody labels the corresponding glandular epithelia.
16.71	9239523	TITLE: Gene expression of zinc-alpha 2-glycoprotein in normal human epidermal and buccal epithelia.
15.25	9328826	Zinc-alpha 2-glycoprotein (Zn alpha 2gp) is almost ubiquitous in body fluids, and its antibody labels the corresponding secretory epithelia.

Ranked Abstracts

The **StAT** Processor uses the calculated Z-score to determine the most significant abstracts. An average Z-score is calculated for each abstract. The abstract level Z-scores is then used to organize the abstracts in order from most significant (most keywords, highest Z-score) to least significant (fewest keywords, lowest Z-score). The keywords within the abstract are highlighted in blue, new words in green, Pharma terms in red and the key sentences are in **bold face**. Links let you move from a highlighted keyword back to the keyword list so that you can review other abstracts containing the same term. If the abstracts were the result of a MEDLINE search, or if user supplied abstracts were in MEDLINE or "FASTA with PMIDs" format, a link to the MEDLINE article allows you to view the abstract in its original form.

StAT Ranked Abstracts

Color Key: new term, pharma term, keyword. (Keywords, Sentences)

>9239523 9.46 (1)

TITLE: Gene expression of zinc-alpha 2-glycoprotein in normal human epidermal and buccal epithelia. Zinc-alpha 2-glycoprotein (Zn alpha 2gp) is almost ubiquitous in body fluids. We have found it to be also present in stratified epithelia. We compare its mRNA expression in cells from human epidermis and buccal mucosa cultured in media of graded differentiation potential (attained by varying calcium ion concentration and adding serum). The Zn alpha 2gp gene is upregulated in both epithelia with differentiation and further with exposure to interferon gamma or transforming growth factor beta 1. **The upregulation by these agents increases with differentiation in epidermal cells, but peaks in the low-differentiation medium in buccal epithelia. We compared gene expression levels of Zn alpha 2gp with those of characteristic cytokeratins of stratified epithelia (k5 for basal cells, K10 for epidermal suprabasal cells, and K13 for mucosal suprabasal cells).** This pattern correlation associates Zn alpha 2gp cell-type dependently with late differentiation, i.e. with keratin K10 in epidermis and with K13 in buccal epithelium.

>9413177 9.19 (2)
TITLE: Modulation by interferon-gamma of zinc-alpha 2-glycoprotein gene expression in human epithelial cell lines. **Zinc-alpha 2-glycoprotein has been detected in most body fluids, and its antibody labels the corresponding glandular epithelia. We have also detected it in human stratified epithelia (epidermis and buccal mucosa).** In this study, the mRNA levels of zinc-alpha 2-glycoprotein were found to be about twice as high in epithelial cells of mucosal origin (whether normal primaries or neoplastic cell lines) as in epidermoid cells (normal epidermal primary cultures, an immortalized but non-tumorigenic epidermal cell line, and neoplastic vulvar and cervical cell lines). **Interferon-gamma strongly upregulated gene expression, but substantially less in mucosal than epidermoid cells.** To compare responses as a clue to the function of zinc-alpha 2-glycoprotein, we ran parallel experiments with three markers of distinct properties, all known to be induced by interferon-gamma. There was the least resemblance for involucrin, a qualitative similarity for HLA-DR, and a rather better match for 2'-5' oligoadenylate synthetase.

Please note, that when the test set is the result of a MEDLINE search, there are instances where an identical abstract will be returned with two or more different identification numbers (PMIDs). In this case, **StAT** displays one copy of the abstract in correct ranked order. Duplicate abstracts are displayed at the end of the list.

Keywords

The **StAT** Processor uses the calculated Z-score and the user entered cutoff to determine the significance of the terms in the set of abstracts. All terms with Z-scores over the cutoff value, and that occur in at least two abstracts, are displayed for the user, ranked in order from highest (most significant) to lowest (least significant) Z-score. For a complete list of terms in the abstract set, you may choose to set the cutoff to zero, though this results in a very large amount of output.

The Z-score for each term is displayed. If the term is not in the background set and it occurs in at least two abstracts, then it is marked as "new". If the term is in the Pharma set then it is marked "pharma". New terms are assigned a Z-value of one-tenth of the maximum Z-score, Pharma terms are set at one-half the maximum Z-score (if Pharma terms are being used). Multiple forms of the Pharma terms are collapsed into a single entry, e.g., inhibit, inhibited, inhibitor becomes "inhibit|ted|tor". In addition, for the first 15 abstracts that contain the term, the abstracts in which each term appears are listed, with links that allow you to jump to the marked abstracts.

StAT Identified Keywords

Color Key: new term, pharma term, keyword. (Sentences, Abstracts)

highZscore = 64.95; new term score = 6.49; pharma term score = 32.47

Keyword	Freq.	Z-Score	Abstract IDs
zag	0.64	new	10208886 11425849 8959578 10607669 9114041 1297243 8409407 7542636 11309332 11205870 9265766 0010846162 9826947 10507755 *etc.* (16 total abstracts)
2-glycoprotein	0.20	new	11425849 9675022 9413177 9239523 9328826
covalent	0.20	new	8959578 11205870 9826947 10507755 11197750
zomepirac	0.20	new	8959578 11205870 9826947 10507755 11197750
glucuronide	0.20	new	8959578 11205870 9826947 10507755 11197750
zp	0.16	new	11205870 9826947 10507755 11197750
epidermi	0.16	new	10698972 9413177 9239523 9328826
zinc-alpha(2)-glycoprotein	0.16	new	11746487 10698972 10462714 11455581
expression	0.40	pharma	10208886 10462714 10698972 11455581 11746487 8409407 9239523 9328826 9413177 9504814
bind|ding	0.40	pharma	10206894 10208886 10507755 10698972 11197750 11425849 7542636 8959578 9114041 9265766
inhibit|ted|tion|tor	0.28	pharma	10462714 11455581 11746487 1297243 7542636 9675022 9826947
expressed	0.20	pharma	10462714 11309332 8409407 9328826 9504814
drug	0.20	pharma	10507755 11197750 11205870 8959578 9826947
crystal|llography	0.16	pharma	0010846162 10206894 11425849 9114041
x-ray	0.08	pharma	11425849 9114041
upregulated	0.08	pharma	9239523 9413177
epithelia	0.32	64.95	11746487 9675022 9413177 9239523 9328826 11309332 10462714 11455581
antiinflammatory	0.16	56.26	8959578 11205870 9826947 11197750
stratified	0.20	52.30	11746487 9413177 9239523 10462714 11455581
streptococcus	0.12	42.98	10208886 1297243 7542636
fluid	0.48	40.80	11425849 9114041 11746487 9504814 9675022 9413177 9239523 9328826 10462714 0010846162 11455581 10206894
acyl	0.20	38.01	8959578 11205870 9826947 10507755 11197750
metabolite	0.20	36.64	8959578 11205870 9826947 10507755 11197750

• Hints and Suggestions

Remember that phrases, multi-word names and other space containing phrases must be quoted; for example "estrogen receptor".

MEDLINE queries containing multiple terms without a boolean search operator will be treated as implied ANDs. In general, if you are searching multiple terms, such as alternate names or pseudonyms for a gene or protein, you need to use an explicit OR between the terms such as;

"zinc alpha-2 glycoprotein" **OR** zag

Typically, it takes about one minute to retrieve and process 500 abstracts and about 2 more minutes to download the entire output.

• Known and Potential Problems

If you limit the abstracts to less than the total available for the query in MEDLINE, then the most recent abstracts are used.

8. Index